

T

Thermal Energy Storage

RAINER TAMME, DOERTE LAING, WOLF-DIETER STEINMANN,
THOMAS BAUER
Institute of Technical Thermodynamics, German
Aerospace Center, Stuttgart, Germany

Article Outline

Glossary
Definition of the Subject
Introduction
Principles of Sensible Heat Storage
Principles of Latent Heat Storage
Principles of Chemical Heat Storage
Thermal Energy Storage for Concentrated Solar Power
Future Directions
Bibliography

Glossary

Latent heat storage Is connected with a phase transformation of the storage materials (phase change materials – PCM), typically changing their physical phase from solid to liquid and vice versa. The phase change is always coupled with the absorption or release of heat and occurs at a constant temperature. Thus, the heat added or released cannot be sensed and appears to be latent. Stored energy is equivalent to the heat (enthalpy) for melting and freezing.

Sensible heat storage Results in an increase or decrease of the storage material temperature stored energy is proportional to the temperature difference of the used materials.

Thermochemical heat storage Is based on reversible thermochemical reactions. The energy is stored in the form of chemical compounds created by an endothermic reaction and it is recovered again by

recombining the compounds in an exothermic reaction. The heat stored and released is equivalent to the heat (enthalpy) of reaction.

Definition of the Subject

Thermal energy storage (TES) is a key element for effective and increased utilization of solar energy in the sectors heating and cooling, process heat, and power generation. Solar thermal energy shows seasonally (summer–winter), daily (day–night), and hourly (clouds) flux variations which does not enable a solar system to provide heat or thermal power according to the demand profile of specific users. Strategies to overcome this problem are to operate in hybrid (solar + fossil) mode and/or to use TES. By integrated TES solar thermal system shows significant benefits:

- Solar contribution is increased, allows operation during times of no solar radiation.
- Rapid flux variations can be compensated (avoid strong gradients for connected components, e.g., piping, heat exchanger, boiler, turbines, etc.) which increase lifetime of components.
- Surplus energy can be used and does not need to be dumped.
- Size of subsequent components, e.g., evaporator, condenser, boiler, turbines, can be reduced.
- Allows improved thermal management of the solar system (e.g., increased start-up time, accurate preheating of solar steam cycle).
- Can be used to cover peak demand.

Introduction

A characteristic of thermal energy storage systems is that they are diversified with respect to temperature, power level, and heat transfer fluids and that each application is characterized by its specific operation parameters. This requires the understanding of a broad portfolio of storage designs, media, and methods.

Heat – in the physical sense – is a form of energy and can be stored in various ways and for many different applications. *Low-temperature heat* is stored for heating, ventilation and air-conditioning (HVAC), and domestic hot water supply, and *high-temperature heat* for industrial processes and solar thermal power plants. Thermal energy storage can be classified according to the heat storage mechanism in sensible heat storage, latent heat storage, and thermochemical heat storage. For the different storage mechanisms, Fig. 1 shows the working temperature and the relation between energy density and maturity.

Each storage concept has its best suited materials and these may occur in different physical phases: as solids, liquids, or via phase change. For example, the volumetric and gravimetric energy densities of the materials have a decisive impact on the capacity of the storage system. The thermal conductivity of the materials is important for the charge and discharge power of the storage system. Furthermore, small density changes versus temperature minimize thermomechanical stress phenomena. A decisive criterion of a heat storage medium is its price and the costs that arise upon its utilization. Long life and a high cycling stability are prerequisites for economic

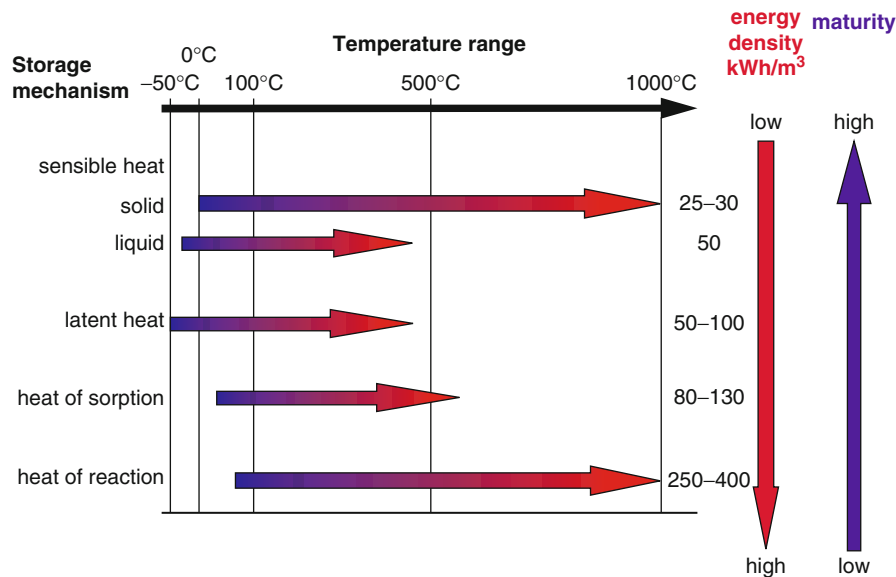
application, i.e., at a price competitive with existing storage facilities.

The charge and discharge powers of sensible and latent heat storage systems are determined mainly by heat transfer processes. For thermochemical storage, mass and heat transfer processes are the dominating physical effects.

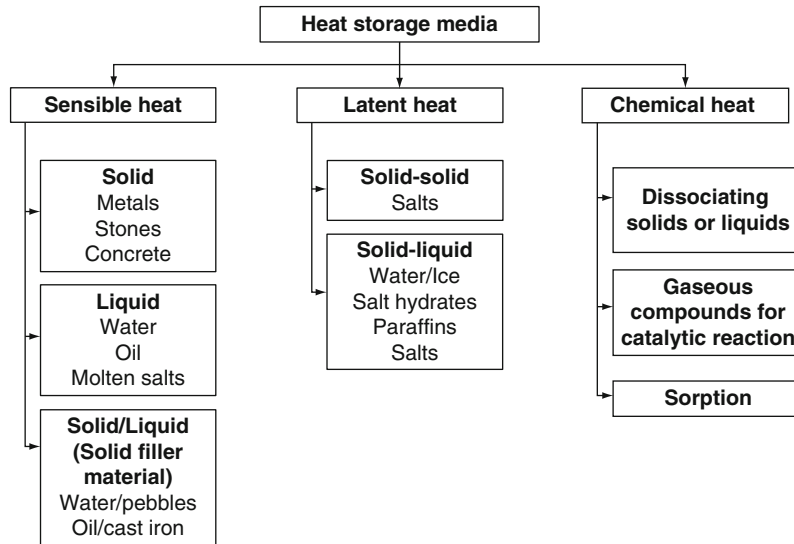
The heat of the carrier fluid may be either transferred *directly* to the storage material as, for example, in a dry pebble bed with air flow, or a heat exchanger may be required as in a solar domestic hot water store where the water–antifreeze mixture flowing through a solar collector has to be separated from the hot water for consumption (*indirect* concept). Other aspects of selecting a heat storage material may be operational advantages in energy supply systems or a larger flexibility in application. A survey of heat storage media is given in Fig. 2.

Principles of Sensible Heat Storage

Sensible heat always results in an increase or decrease of the material temperature. All materials have a capability of absorbing and storing heat due to the fact that they



Thermal Energy Storage. Figure 1 Thermal energy storage mechanism, its working temperature, and correlation to energy density and status of technical maturity



Thermal Energy Storage. Figure 2
Overview of heat storage media with examples

have a mass m and a specific heat capacity c_p at constant pressure. For a temperature difference $\Delta T = T_2 - T_1$ this heat (or enthalpy) amounts to

$$Q_{\text{sensible}} = m \cdot c_p \cdot (T_2 - T_1) \quad (1)$$

T_2 denotes the material temperature at the end of the heat absorbing (charging) process and T_1 at the beginning of this process. This heat is released in the respective discharging process.

In Table 1 some characteristic materials are listed together with their thermophysical properties. It needs to be considered, that some material values, such as graphite, are strongly temperature dependent. Also, for example, impurities in metals cause a drop in the thermal conductivity values. Equation 2 defines the thermal diffusivity a , where λ is the thermal conductivity and ρ the density. Equation 3 gives the definition of the heat diffusivity b .

$$a = \frac{\lambda}{\rho \cdot c_p} \quad (2)$$

$$b = \sqrt{\lambda \cdot \rho \cdot c_p} \quad (3)$$

Metals and graphite are best suited for quick charging and discharging (high thermal diffusivity a) and for a large amount of heat stored in a given time (high heat diffusivity b). Other solid materials such as stones

are much less advantageous; their respective values are smaller by an order of magnitude.

Liquids offer the advantage of a possible use as both storage medium and heat transfer fluid. The most widely applied media, in this respect, are water and thermal oil.

Solid Storage Materials

Solid materials can be utilized in a wide temperature range and heated up to very high temperature (e.g., refractory bricks in Cowper regenerators to 1,000°C). Solids are often chemically inert and have a low vapor pressure. In addition, the containment can often be simpler compared to systems based on liquids. Solid storage materials can be classified as metals and nonmetals.

Natural materials in the form of rocks and pebbles are abundant and cheap. For low temperatures, rock and soil can be used as ground storage. For high temperatures, the thermomechanical stability of the solids is important. Here, rocks such as granite, basalt, and quartzite, as well as pebbles can have suitable properties.

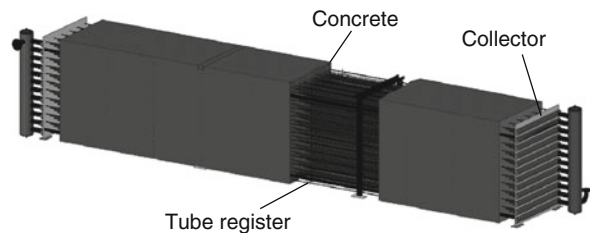
Of the manufactured solid materials, various ceramics are widely used as heat storage materials. In the low-temperature range, bricks act as a buffer

Thermal Energy Storage. Table 1 Thermophysical properties of media for sensible heat storage

Material	$T, ^\circ\text{C}$	$\rho, (\text{kg}/\text{m}^3)$	$c_p, (\text{kJ}/[\text{kg}\cdot\text{K}])$	$\lambda, (\text{W}/[\text{m}\cdot\text{K}])$	$10^6 \times a, (\text{m}^2/\text{s})$	$10^{-3} \times b, (\text{J}/[\text{m}^2\text{Ks}^{1/2}])$
Solids						
Aluminum 99.99%	20	2,700	0.945	238.4	93.3	24.66
Copper (commercial)	20	8,300	0.419	372	107	35.97
Iron	20	7,850	0.465	59.3	16.3	14.7
Lead	20	11,340	0.131	35.25	23.6	7.24
Brick (dry)	20	1,800	0.84	0.50	0.33	0.87
Concrete (aggregates)	20	2,200	0.72	1.45	0.94	1.52
Granite	20	2,750	0.89	2.9	1.18	2.67
Graphite	20	2,200	0.61	155	120	14.41
Limestone	20	2,500	0.74	2.2	1.19	2.02
Sandstone	20	2,200	0.71	1.8	1.15	1.68
Slag	20	2,700	0.84	0.57	0.25	1.13
Sodium chloride	20	2,165	0.86	6.5	3.5	3.5
Soil (clay)	20	1,450	0.88	1.28	1.0	1.28
Soil (gravelly)	20	2,040	1.84	0.59	0.16	1.49
Liquids						
Molten salt (K-NaNO ₃)	230	1,950	1.57	0.50	0.16	1.24
Paraffin (liquid)	20	900	2.13	0.26	0.14	0.71
Silicone oil (AK250)	25	970	1.465	0.168	0.118	0.49
Sodium	100	927	1.385	85.84	66.85	10.50
Transformer oil	60	842	2.09	0.122	0.069	0.46
Water	20	998	4.183	0.598	0.142	1.58

for the climatization of buildings. At higher temperatures, refractory bricks based on oxides (silica, alumina, magnesia, and iron oxide – feolite), carbonates (e.g., magnesite), and their mixtures are commercially utilized in applications such as Cowper regenerators, night-storage heaters, and tiled stoves.

For high temperatures, concrete is another attractive sensible heat storage material, because of its low cost due to inexpensive aggregates, the availability throughout the world and the available and simple processing route via casting. Figure 3 shows a concrete storage unit which can operate up to 400°C [1, 2].

**Thermal Energy Storage. Figure 3**

Schematic layout of a concrete storage module showing the main elements: tube register, concrete storage material, and heat transfer fluid collector

Typically, metals exhibit higher thermal conductivities but are less attractive from an economic point of view compared to nonmetals. This high conductivity is advantageous in terms of a fast charge and discharge process of the storage system. High conductivity metals, such as aluminum and copper, can be used in electronics as a heat sink and for thermal management. *Cast iron* is an alternative material with a high thermal conductivity and volumetric heat capacity (energy density). It is used for high-temperature storage together with oil as heat carrier.

The heat transfer concept of storage systems using solid materials is usually based on an additional fluid as a heat carrier (e.g., water, steam, air, oil, molten salt) for the charge and discharge process. The heat carrier fluid may be operated either in direct contact (e.g., Cowper regenerators in the high-temperature glass and steel industry) or indirect contact (e.g., via a tubular heat exchanger) with the solid storage material.

Direct contact design is realized as a packed bed of solid particles in a container, such as a pebble bed or a rock pile, or in a more defined arrangement such as a regular array of checker bricks. Heat has to be carried to or from such stores by a heat carrier such as air or water. These beds provide a large surface area for heat transfer in their dispersed materials; the internal heat losses are only small because the particles have little surface contact to surrounding particles and thus, little conduction losses. For low temperatures, reduced insulation around the storage container is sufficient when air – with a small thermal conductivity – is used as a heat carrier and fills the gaps between the particles. Such particle beds, however, cause a large pressure drop in the fluid flow. This has to be compensated by pumps or compressors. Another drawback is the large storage volume required. Conveniently, this is provided in underground installations.

As the solid storage materials are in direct contact with the fluid heat carriers, properties of the solids have to be taken into account which affect fluid flow and heat transfer, such as particle and container size, mechanical durability against abrasion, and thermal cycling and hygroscopy.

The thermal stratification (see section “[Liquid Storage Materials](#)”) is well maintained in a heat store,

because natural convection in the heat carrier is suppressed and internal heat conduction is low, because of the aforementioned reasons. The combination of hollow bricks and air has been known as hypocaust-heating systems since antiquity. Water-flooded pebble beds are known as artificial aquifers.

In some cases the direct contact of heat carrier and storage material is not economic or feasible, for example, if the heat carrier is pressurized (e.g., steam) or incompatible with the storage medium. In this case, a design with indirect contact of carrier and storage medium can be utilized. The previous shown example of a high-temperature concrete storage using a steel tube register with pressurized water, steam, or oil as heat carrier ([Fig. 3](#)) represents this case [[1](#), [2](#)].

Liquid Storage Materials

The most widely used liquid for thermal storage is water. Water has the following advantages:

1. It is abundant and inexpensive.
2. It has a relatively high heat diffusivity ($b = 1.58 \times 10^3 \text{ J m}^{-2} \text{ K}^{-1} \text{ s}^{-1/2}$) and a relatively low thermal (temperature) diffusivity ($a = 0.142 \times 10^{-6} \text{ m}^2/\text{s}$), which is an advantage for thermal stratification within a hot-water storage tank.
3. It can be easily stored in all kinds of containers.
4. The control of water flow systems is highly flexible and is often state of the art.
5. Water can be directly used without heat exchangers; it is easy to handle, nontoxic, and noncombustible.
6. It is easily mixable with additives (antifreeze, anticorrosive).

Water has also some disadvantages, which can be summarized as follows:

1. It only provides a limited operation range between 0°C (freezing) and 100°C (boiling).
2. It is corrosive.
3. It has high vapor pressures ($p > 5 \text{ bar}$) at temperatures $>155^\circ\text{C}$.

The vapor pressure can be reduced when mixtures of water with other chemicals are applied; e.g., a mixture of 50% H_2O and 50% NaOH can be used

up to 140°C without a pressure vessel. The liquid material storage approach can be realized as a single tank or a two-tank concept consisting of two individual tanks at different temperature and fluid level. In a single tank, hot and cold media require thermal stratification to reduce losses.

Thermal stratification in heat storage vessels is desired, because the value of high temperature heat is maintained in the upper part of the vessel, while low-temperature fluid (as backflow from a heat consumer) can still be stored in the lower part of the vessel. The temperature distribution within a storage vessel varies with time. Due to transient internal heat conduction, the stratification is destroyed in the course of time, even in very well insulated vessels.

Applications of Water Storages for Solar Energy

Storage tanks for hot water are used in industry and dwellings. They come in sizes of 0.1 m³ (domestic hot water storage) to 12,000 m³ (solar-assisted district heating). The most recent development is “solar combination stores” of about 0.5 m³ which serve for both domestic hot water supply and house heating.

Water is commonly used also in large thermal stores, both as storage medium and heat transfer fluid. In large stores (e.g., seasonal stores for solar energy), the storage capacity can be turned over only once or – at the most – three times a year. Such a small energy turnover makes the stored heat expensive unless the storage container is inexpensive.

A technical/economical boundary for *steel tanks* is a size of around 100,000 m³. For *ground containers*, the waterproof liner may create problems (e.g., tightness after welding, perforation by stones): plastic liners and stainless steel liners are used. These large containers are built of concrete and they are fully or partly buried in the ground. Their capacity is approximately 80 kWh/m³. In summer, hot water is fed to the water layers on top and cold water withdrawn from the bottom; in winter, hot water is withdrawn from the top layers and cold water referred at the bottom.

In so-called *ground stores* low-temperature heat is stored in soil or pebbles. The storage medium in such cases is a mixture of water and solid particles. Density ρ and specific heat capacity c of these mixtures can be obtained from the mixing laws. The thermal conductivity λ depends strongly on the composition of the mixture.

Although some mathematical equations exist in the literature, reliable data for λ depend on measurements. Heat is usually exchanged with the ground by tubes containing a circulating carrier fluid, mostly water. These tubes may be placed close to the surface (0.5–1 m below) in a meandering arrangement or spirally wound in horizontal ditches. For large stores (more than 10,000 m³ water equivalent storage capacity), vertical holes are drilled for the heat exchanger tubes.

In some areas, geology allows for a so-called *aquifer store*. This requires a permeable layer of soil (pebbles or sand) enclosed between horizontal layers of impermeable material (clay). Under such advantageous conditions, two fountains can be drilled some distance apart from each other and hot water can be inserted in summer and withdrawn in winter. Problems can arise from minerals dissolving in the water and precipitating on tubes and heat exchangers.

So-called *man-made aquifers* are built in sizes of 10,000 m³; the ground excavations are filled with pebbles and water and sealed by plastic sheets. In all cases where mixtures of water and solid material are used for heat storage, the storage volume has to be larger than in the case of pure water. Rough data for the different types are given in Table 2.

Large-scale solar thermal storage in water is possible in *solar ponds*. These ponds act both as solar collector and as storage. In this concept the water itself is used as an insulator. The convection in the water can be suppressed by different methods. One common method is to utilize a density gradient which corresponds to a salt concentration gradient.

Molten Salt For temperatures above 100°C, molten salts are attractive candidates for sensible heat storage.

Thermal Energy Storage. Table 2 Storage volume of different storage media compared to pure water for 1 m water equivalent and 60 kWh capacity at $\Delta T = 50$ K

Storage media	Storage volume
Hot water storage	1 m ³
Pebble – water storage	1.3–2 m ³
Aquifer storage	2–3 m ³
Borehole storage	4–5 m ³

The major advantages of molten salts are low costs, high thermal stabilities, and low vapor pressures. The low vapor pressure results in storage designs without pressurized vessels. Also, many salts can be operated in air contact without significant degradation. In general there is experience with molten salts from a number of industrial applications related to heat treatment, thermochemical reactions, and heat transfer. For solar thermal power plants, large-scale two-tank storage systems with a capacity of several ten thousand tons have been built (see section “[Indirect Storage in Liquid Media](#)”). Typically, a non-eutectic salt mixture of 60 wt.% sodium nitrate and 40 wt.% potassium nitrate is utilized. The eutectic mixture has a melting temperature of 222°C and the thermal stability limit is at about 550°C. One major difficulty with molten salts is unwanted freezing during operation. Hence, often additional heating systems are required.

Further Liquid Storage Materials Oil is a common heat transfer fluid and has also been used as a liquid storage material. For example, mineral oil can be used at ambient pressures up to about 300°C. Synthetic oils are thermally stable up to around 400°C, but at higher temperatures they have to be pressurized which is often uneconomic.

For media other than water, the liquid storage material often creates the predominate cost of the entire storage system. In these cases low-cost solid filler material can replace the more expensive liquid storage material (e.g., cast iron in oil, molten salt thermocline designs).

Principles of Latent Heat Storage

Materials for the storage of latent heat are also named phase change materials (PCM), because of the fact that they change their physical phase from solid to liquid and vice versa [3–13]. The phase change is always coupled with the absorption of heat when the solid melts – heat of melting – and a heat release when the liquid solidifies. The phase change occurs at the melting temperature T_m . At that temperature, these materials melt when heat is added, but do not exhibit a rise in temperature. Thus, the heat added cannot be sensed and appears to be latent.

A change of one crystalline form into another without a physical phase change may also be

considered as storage of latent heat. The phase change of salt hydrates is considered a solid–liquid transformation. The solid hydrate crystals release their water of crystallization at a certain temperature (or a temperature range) and the anhydrous salt dissolves in this water. This process is reversed when heat is withdrawn. The heat (or enthalpy at $p = \text{const.}$) stored with a phase change amounts to

$$q_{\text{latent}} = \Delta h_m \quad (4)$$

where Δh_m is the latent heat (enthalpy) of melting. When a material is heated from an initial temperature T_1 to the melting temperature T_m and the melt is further heated to T_2 , the total heat stored is

$$Q_{\text{total}} = Q_{\text{sensible,solid}} + Q_{\text{latent}} + Q_{\text{sensible,liquid}} \quad (5)$$

$$= q_{\text{total}} \cdot m$$

or using (1), (4), and (5):

$$Q_{\text{total}} = m \cdot c_{p,\text{solid}}(T_m - T_1) + m \cdot \Delta h_m \quad (6)$$

$$+ m \cdot c_{p,\text{liquid}}(T_2 - T_m)$$

The specific heat capacity of the solid phase is $c_{p,\text{solid}}$ and of the liquid phase $c_{p,\text{liquid}}$.

The heat of fusion (or heat of melting) is much greater than the sensible heat. Therefore, phase change materials have a larger volumetric energy storage capacity than sensible storage materials, if smaller temperature intervals are considered. A further advantage of PCMs is the absorption and the release of heat at a constant temperature; this melting temperature is material-specific and the material can be selected accordingly.

In most cases the heat of the solid–liquid phase change is utilized. The heat of solid–solid phase changes (rearrangement of crystalline forms) is typical smaller compared to the heat of melting and solidification. Although enthalpies are large, the liquid–gas phase change is not utilized for latent heat storage due to the large volume of the gas phase.

Latent heat storage materials should be inexpensive and be characterized by:

1. A large phase-change enthalpy and a high density
2. A large thermal diffusivity in the solid and liquid phase
3. Thermal stability and cycle stability with a low vapor pressure

4. Little or no subcooling during freezing and little or no supersaturation during melting
5. A low thermal expansion, i.e., low volume change during melting

They should neither exhibit segregation (e.g., like Glauber's salt), nor be poisonous, flammable, or form explosive phases. They should not be corrosive, or react with materials in contact, such as containment, heat

exchanger, and heat transfer enhancement structure. Various phase change materials are listed together with their thermophysical properties in [Table 3](#). Thermophysical properties are not always available and their values may differ among different literature sources. Impurities in the substances can change the properties considerably. The table presents only characteristic materials; many more are found in literature.

Thermal Energy Storage. Table 3 Thermophysical properties of solid to liquid phase-change materials. The values are related to the solid (s) and liquid (l) phase below and above the respective melting point

Salt system	T (°C)	H (J/g)	λ_s (W/mK)	λ_l (W/mK)	$C_{p,s}$ (J/gK)	$C_{p,l}$ (J/gK)	ρ_l (g/cm ³)	$\Delta V/V_s$ (%)
Ice and salt hydrates								
H ₂ O	0	334	2.3	0.56	2.10	4.22	1.00	−8.3
KF·4H ₂ O	19	230	n.a.	n.a.	1.8	2.4	1.46	n.a.
CaCl ₂ ·6H ₂ O	30	180	1.1	0.5	1.4	2.1–2.3	1.56	13
Na ₂ SO ₄ ·10H ₂ O	32	250	0.5	n.a.	1.9	3.3	1.41	4
Na ₂ HPO ₄ ·12H ₂ O	35	280	0.5	0.5	1.5–1.7	2.0–3.2	1.44	5
Na(CH ₃ COO)·3H ₂ O	58	250	0.7	0.4	2.1	3.0	1.28	13
Pure metals								
Na	98	113	122	87	1.36	1.38	0.93	2.8
Zn	419	112	100	50	0.45	0.48	6.76	1.9
Al	660	397	211	91	1.18	1.18	2.38	7.5
Anhydrous salts and their mixtures (Composition in weight%)								
KNO ₃ –LiNO ₃ (67–33)	133	170	n.a.	n.a.	n.a.	n.a.	1.9	14
KNO ₃ –NaNO ₂ –NaNO ₃ (53–40–7)	142	80	0.5	0.5	1.3	1.6	1.98	4
LiNO ₃ –NaNO ₃ (49–51)	194	265	n.a.	0.5	n.a.	n.a.	1.9	13
KNO ₃ –NaNO ₃ (54–46)	222	100	n.a.	0.5	1.4	1.5	1.95	5
LiNO ₃	254	360	1.4	0.6	1.8	1.6–2.0	1.78	22
NaNO ₂	270	180	0.7–1.3	0.5–0.7	n.a.	1.6–1.8	1.81	17
NaNO ₃	306	175	0.6	0.51	1.66	1.66	1.91	11
NaOH	322	210	0.9	0.8	2.0	2.1	2.13	16
KNO ₃	337	100	n.a.	0.4–0.5	1.4	1.3–1.4	1.87–1.89	3
NaCl–KCl–MgCl ₂ (24.5–20.5–55)	393	240	1.0	n.a.	1.0	1.0	1.8	n.a.
K ₂ CO ₃ –Li ₂ CO ₃ –Na ₂ CO ₃ (35–32–33)	397	275	n.a.	n.a.	1.7	1.6	1.90	17
K ₂ CO ₃ –Li ₂ CO ₃ (65–35)	505	345	n.a.	n.a.	1.3–1.8	1.8–2.3	1.96	10

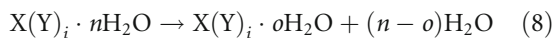
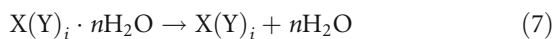
Cold Storage (0°C and Below)

For refrigeration below 0°C, ice slurries using a solution of water–salt (brine) and water–glycol can be utilized. Applications are industrial processes and space cooling. Within the solution fine solid crystals are suspended. A major advantage of the slurries is that they can be pumped. Hence, they can act as both the heat carrier and the heat storage medium. The slurry has the advantage of a larger heat capacity compared to the solution.

For space cooling, latent heat water–ice storage systems are commercially available. These systems utilize different heat transfer concepts. In the ice-on-coil design the ice grows around the heat exchanger tubes. Another concept utilizes spherical capsules filled with the PCM water/ice (macroencapsulation). These capsules are surrounded by the heat transfer fluid. Both, coils and capsules are commonly made of polymers such as polyethylene. At this stage, alternative direct contact heat transfer concepts to reduce costs of the heat exchanger or the capsules are examined but have not evolved commercially. Another very old concept is the use of natural ice (or snow) as a seasonal storage to supply cold in summer. Nowadays, there is an interest in the cooling of buildings and many snow storage demonstration plants have been built. For the transportation of food, beverage, and other products, smaller insulated boxes with encapsulated ice modules are commonly used.

Low-Temperature Storage (0–120°C)

Salt Hydrates Many salt hydrates are sufficiently inexpensive; they are only little corrosive and slightly toxic. They exhibit only small changes in volume during phase change and have a relatively high heat of fusion (compared to organic compounds) at low melting temperature. Salt hydrates are of the chemical form $X(Y)_i \cdot n \text{H}_2\text{O}$ with $X(Y)_i$ representing an inorganic compound. When the salt hydrate is heated to its transition temperature, one of the following reactions will take place:



At the melting point, the hydrate crystals split up into either anhydrous salt and water (7) or into

a hydrate with a lower number of water molecules and water (8). Several salt hydrates have two major drawbacks for thermal storage: They suffer from incongruent melting and supercooling.

Incongruent melting occurs when not enough water of crystallization is released during thawing and the anhydrous salt cannot completely dissolve. The solid residual part settles at the bottom of the container due to its higher density. In the reverse process (during recrystallization), some fraction of the precipitated solid is unable to come into contact with the water necessary for its recrystallization. Thus, the useful heat of fusion reduces after several charge–discharge cycles. Incongruent melting can be prevented by:

1. Addition of a thickening agent which holds the anhydrous solid part in suspension and prevents its settling
2. Stirring, vibrating, or rotating the material
3. Addition of excess water
4. Encapsulation of the material (e.g., in plastics, as metallic containers may be corroded)

Supercooling occurs during cooling (heat discharge) when no sufficient nucleation sites are available. The crystallization does not take place at the melting point; the liquid is cooled below that point and the stored heat is not discharged at the expected melting temperature, but at some arbitrary temperature. In some cases it might happen that the stored heat cannot be withdrawn, because the temperature difference between the subcooled melt and the required fluid temperature is not high enough. Supercooling can be prevented by addition of small quantities of a nucleating agent with a crystal structure similar to that of the salt.

Paraffins Paraffins are waxy solids at room temperature. Their melting points and heat of fusion increase with molar mass. Compounds in the series $\text{C}_{14}\text{H}_{30}$ – $\text{C}_{40}\text{H}_{82}$ have melting points in the range 6–80°C. They are advantageous storage materials for the following reasons:

1. They are widely available and inexpensive.
2. Noncorrosive and nontoxic.
3. Chemically stable, with a low density.
4. Phase transformation is rapid.
5. Supercooling is negligible.

But they show the following drawbacks:

1. Low volumetric energy density (compared to salt hydrates).
2. Change in volume during melting by about 10%.
3. Low thermal conductivity and diffusivity.
4. No distinct melting point, because cheap technical paraffins used for thermal storage are mixtures of many compounds.
5. Not compatible with plastics.
6. They are flammable.

Applications for Solar Heating and Cooling

There are several commercial PCM products without a heat carrier fluid (passive systems) [10]. In these applications the PCM increases the heat capacity of the system and the PCM stabilizes the temperature of the system. Commercial PCM products are available in the areas of insulated transport containers (e.g., medical applications), the thermal management of electronic equipment, electric heating systems (e.g., floor heating), and human body comfort (e.g., pocket heater, clothes). Another area of commercial products is the space cooling of buildings. In particular in lightweight buildings with a low thermal mass, the PCM can reduce temperature fluctuation and cut peak temperatures. There are different ways to insert the PCM in the building. They include the additional use of boards and panels, such as gypsum plasterboard with microencapsulated paraffin and aluminum bags filled with salt hydrates. Another option is the integration of the PCM into the building material, such as plaster and concrete containing microencapsulated PCM.

For the heating and cooling of buildings, PCM systems with a heat carrier fluid, commonly water and air, have been researched [10]. The advantage of these active systems is that they have a higher heat transfer rate (or shorter charge/discharge time) compared to passive systems. For buildings, the discharge of the PCM at night is a key issue. Cooling of the PCM in the building can be achieved by a cold forced convection stream of night air which passes the PCM unit. The PCM systems can be located in the floor, ceiling, wall, or ventilation channels. In some cases the outside air is too warm at night and cooling with night air cannot be utilized. In this situation, water or brine as a heat carrier fluid and alternative cold sources can be used. The cold can be from natural sources

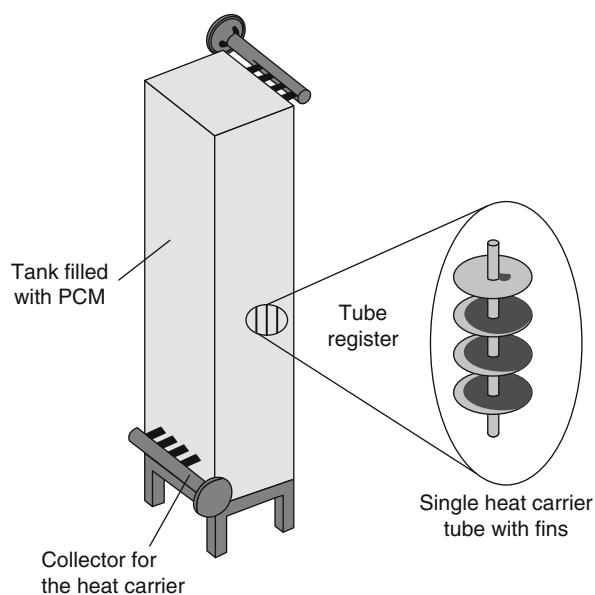
(e.g., groundwater and soil) or from artificial sources (e.g., absorption chiller). Typically, the water or brine flows through capillary tubes within the PCM panels or boards. The replacement of conventional hot water storage systems, mainly for space heating and tap water in households, is also of research interest. Typical storage temperatures range from 40°C to 70°C. The heat transfer is usually limited by the low thermal conductivity of the PCMs. Heat transfer designs include mainly indirect contact concepts (e.g., heat exchanger tubes with and without fins, as well as encapsulated-PCM in the water), but also other concepts (direct contact concepts, PCM slurries).

High-Temperature PCM Storage (Above 120°C)

For the temperature range above 120°C, organic PCMs have been proposed. Their critical aspects have been not fully assessed. They include the long-term thermal stability, the reactivity with air oxygen, and the high vapor pressure. Some disadvantages can be overcome if hermetically sealed storage systems are utilized.

The thermal stability of inorganic materials is inherently higher. For temperatures from 120°C to 1,000°C, inorganic anhydrous salts can be utilized. There are solid–solid and solid–liquid phase transitions of anhydrous salts which can be used. The cations are mainly alkali (e.g., Li, Na, K) and alkaline earth (e.g., Ca, Mg) metals. Anions which are considered include nitrates, nitrites, hydroxides, bromides, carbonates, chlorides, sulfates, and fluorides. Many anhydrous salts are miscible and this results in a large variety of potential single salts and salt mixtures (binary and ternary systems). Table 3 shows selected examples of PCMs. For temperatures up to 350°C, in particular alkali metal nitrates and nitrites and their mixtures are suitable PCMs, if requirements such as handling and steel compatibility are taken into account [11].

In the temperature range 120–320°C, steam storage systems based on PCMs have been developed. These systems utilize a fine parallel pipe heat exchanger which is integrated into the PCM (Fig. 4). During the charging process, the steam condenses inside the pipes and the PCM melts. During discharging, the solidifying PCM releases heat, the water evaporates inside the pipes, and steam is generated. Applications of these



Thermal Energy Storage. Figure 4
Schematic drawing of a PCM storage module with integrated fins for enhanced heat transfer

steam storage systems include the areas of solar thermal power generation, industrial process heat utilization, and combined heat and power generation [12, 13].

Heat Transfer Concepts and Composites

The thermal conductivity of the PCM affects significantly the power density and the heat transfer design of the system. The power density of the storage systems corresponds to the melting and solidification time of the PCM. A typical design is a tubular heat exchanger integrated into the storage material and many applications require melting and solidification times in the order of hours. As can be seen in Table 3, the thermal conductivity of all common PCMs are lower than 1 W/(m.K). These conditions result in a storage design with a large number of heat exchanger tubes and this design is often not economic. Hence, cost-effective storage systems require the development of concepts that compensate this low thermal heat conductivity.

A number of heat transfer enhancement concepts has been proposed [12, 13]. The thermal conductivity can be enhanced by using composite materials, where the properties of a high latent heat of a PCM and a good thermal conductivity of an additive are combined. An

aspect of interest is the form stability of the composites without phase separation after several melting and solidification cycles. For low-temperature applications ($<100^{\circ}\text{C}$), paraffin/graphite composites have been successfully demonstrated. Other approaches aim for an extended heat transfer area between the heat carrier and the PCM. This can be achieved by the encapsulation of the PCM. The capsules provide a large heat transfer area for the surrounding heat carrier fluid (e.g., water or air). Another demonstrated approach is the utilization of finned heat carrier tubes to increase heat transfer. Other conductive structures inserted in the PCM have been also examined.

Principles of Chemical Heat Storage

Reversible thermochemical reactions can be used for storing thermal energy [9, 14]. The energy is stored in the form of chemical compounds created by an endothermic reaction and it is recovered again by recombining the compounds in an exothermic reaction. The heat stored and released is equivalent to the heat (enthalpy) of reaction.

The enthalpy of reaction, generally, is much larger than the enthalpy of transition (in latent heat storage) or the sensible heat stored over a reasonable temperature span. The storage density (based on solid mass or volume) is much larger for thermochemical storage materials than for latent or sensible heat storage materials. Table 4 shows materials for the storage of chemical heat, their temperature of reaction, and their heat of reaction related to the educt reactant. The determination of the volumetric energy density of the system Δh_{sys} shows that storage of gaseous reaction products at 50 bar results in a low Δh_{sys} compared to the volumetric energy density of the educt $\Delta h_{r,\text{educt}}$. On the other hand, storage of liquid (or liquefied) reaction products or utilization of atmospheric gases (e.g., O_2) leads to a high volumetric energy density of the system Δh_{sys} .

The reversible reactions suitable for heat storage can be classified into thermal dissociation reactions and catalytic reactions.

Gas–Solid Dissociation Reactions

Certain solid or liquid compounds can undergo dissociation reactions when they are heated. A gas is released

Thermal Energy Storage. Table 4 Temperature and heat of reaction of materials for thermochemical heat storage

Material	T_r (1 bar) (°C)	$\Delta h_{r,educt}$ (kJ/mol)	$\Delta h_{r,educt}$ (KWh/t)	$\Delta h_{r,educt}^a$ (KWh/m ³)	$\Delta h_{sys}^{a,b}$ (KWh/m ³)
Hydroxides					
$Mg(OH)_2 \leftrightarrow MgO + H_2O$	268	78	372	442	323
$Ca(OH)_2 \leftrightarrow CaO + H_2O$	521	100	373	410	324
$Ba(OH)_2 \leftrightarrow BaO + H_2O$					
Ammonium salt					
$NH_4HSO_4 \leftrightarrow NH_3 + H_2 + SO_3$	467	337	813	727	136
Salt hydrates					
$MgSO_4 \cdot 7H_2O \leftrightarrow MgSO_4 + 7H_2O$	122	411	463	389	272
$CaCl_2 \cdot 2H_2O \leftrightarrow CaCl_2 \cdot H_2O + H_2O$	174	48	91	84	75
$CuSO_4 \cdot 5H_2O \leftrightarrow CuSO_4 \cdot H_2O + 4H_2O$	104	226	251	287	216
$CuSO_4 \cdot H_2O \leftrightarrow CuSO_4 + H_2O$	205	73	114	163	142
Peroxide salts					
$BaO_2 \leftrightarrow BaO + \frac{1}{2}O_2$	782	75	123	305	66
$KO_2 \leftrightarrow \frac{1}{2}K_2O + \frac{3}{4}O_2$	668	101	395	423	64
Carbonates					
$CaCO_3 \leftrightarrow CaO + CO_2$	896	167	463	626	113
$BaCO_3 \leftrightarrow BaO + CO_2$	1,497	212	298	661	139
Metal hydrides					
$MgH_2 \leftrightarrow Mg + H_2$	293	79	834	605	41
$Mg_2NiH_4 \leftrightarrow Mg_2Ni + 2H_2$	253	128	319	–	–
Catalytic reactions					
$NH_3 \leftrightarrow \frac{1}{2}N_2 + \frac{3}{2}H_2$	195	49	799	546	21
$SO_3 \leftrightarrow SO_2 + \frac{1}{2}O_2$	767	98	340	678	82
Reforming					
$CH_4 + H_2O \leftrightarrow CO + 3H_2$	687	205	1,672	n/a	23
$CH_4 + CO_2 \leftrightarrow 2CO + 2H_2$	687	247	1,143	n/a	24

The pressure vessel and other system components are not included

^aSolids: assumed is a powder with 50% pore volume.

Liquids (stored below 50 bar): H_2O , NH_3 , SO_3 , SO_2 .

Gases stored at 50 bar: H_2 , O_2 , CO_2 , O_2 , CO

^bThe volumetric energy density of the system Δh_{sys} is based on the volume of the reactant and products.

while the depleted solid or liquid remains in the reactor (endothermic reaction, i.e., *charging of the store*). The reverse reaction will occur spontaneously if the equilibrium is changed by a temperature decrease

or a pressure increase. Therefore, the dissociation products have to be separated and stored individually. For *discharge*, in the exothermic reaction, the gas is recombined with the solid or liquid.

In general there are various types of gas–solid reaction systems that can be used for thermochemical energy storage. Among them are:

- Dehydration of metal salt hydrates (application in the range of 40–260°C)
- Dehydrogenation of metal hydrides (application in the range of 80–400°C)
- Dehydration of metal hydroxides (application in the range of 250–800°C)
- Decarboxylation of metal carbonates (application in the range of 100–950°C)
- Thermal deoxygenation of metal oxides (application in the range of 600–1,000°C)

The dissociation of calcium hydroxide ($\text{Ca(OH)}_2 \leftrightarrow \text{CaO} + \text{H}_2\text{O}$) has been investigated and reversibility of the reaction could be proven with a reaction enthalpy of 100 kJ/mol and an equilibrium temperature of 521°C at 1 bar. DLR is investigating this reaction system since several years, as especially its temperature range as well as the low material costs and its availability show promising potential for later large-scale application in solar trough power plants [15]. Complete reversibility as well as cycling stability could be shown by thermal analysis.

Concerning central air receivers, a recent project led by General Atomics in cooperation with DLR explores the experimental feasibility of using different multivalent solid oxides for high-temperature thermochemical heat storage [16]. One of the identified promising candidates for such a thermochemical storage system is based on manganese oxide redox reactions [17, 18]: $6 \text{Mn}_2\text{O}_3 \leftrightarrow 4 \text{Mn}_3\text{O}_4 + \text{O}_2$. This step is thermodynamically favorable at relevant temperatures for application with a central air receiver and has a reaction enthalpy of 31.8 kJ per mol of Mn_2O_3 .

Catalytic Reactions

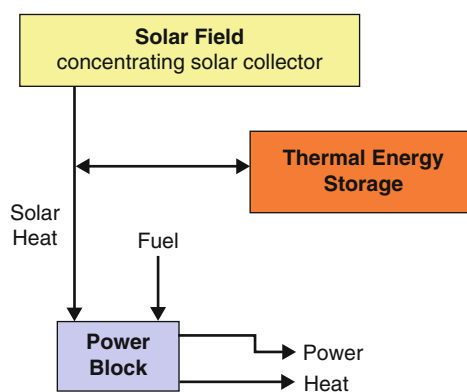
These are usually gas – gas reactions. A catalyst is required to obtain high reaction rates. The reaction products are transported in a pipeline from the charging source to the consumer (discharging sink). This system is called chemical heat pipeline or chemical heat pipe.

Thermochemical storage by heterogeneously catalyzed gas phase reactions has been demonstrated for the

catalytic dissociation of ammonia in a 15 kW solar reactor by the Australian National University where solar heat is used to dissociate NH_3 into nitrogen and hydrogen [19]. The stored heat is retrieved by recombination of N_2 and H_2 to form ammonia. A closed thermochemical cycle for solar-driven methane reforming was successfully tested in a directly irradiated solar reformer in 280 kW scale in a joint project between the Weizmann Institute in Israel and DLR [20]. The main drawback of the approach of heterogeneously catalyzed gas phase reactions is the requirement to separate, compress, and store gases thus adding to the cost and complexity of the process.

Thermal Energy Storage for Concentrated Solar Power

Regarding electric grid and quality of bulk power supply, it is the ability to provide dispatch on demand that makes solar thermal power stand out from other renewable energy technologies like PV or wind. Thermal energy storage systems store excess thermal heat collected by the solar field (Fig. 5). Storage systems, alone or in combination with some fossil fuel backup, keep the plant running under full-load conditions. Improvements on operational flexibility and energy dispatchability by thermal storage and hybridization are identified as key technology objectives for R&D development. The capability of storing high-temperature thermal energy leads to economically



Thermal Energy Storage. Figure 5

Basic layout of a solar thermal power plant with integrated thermal energy storage

competitive design options, since only the solar part of the plant has to be oversized. This solar thermal power plant feature is tremendously relevant, since penetration of solar energy into the bulk electricity market is possible only when substitution of intermediate-load power plants of about 4,000–5,000 h/year is achieved.

From a technical point of view, the storage must have high energy density, good heat transfer between the heat transfer fluid (HTF) and the storage medium, mechanically and chemically stable storage media, compatibility between the heat exchanger, heat transfer fluid and storage medium, complete reversibility, and minimum thermal losses. In a TES system, heat can be stored in the form of sensible, latent, or thermochemical energy.

The option to integrate cost-effective storage systems directly into the facilities represents a significant advantage of solar thermal power plants over other concepts using renewable energy sources. Storage capacity allows not only demand-oriented electricity generation, by compensating fluctuation in solar insolation caused by cloud passing. Storage units facilitate the control of systems using concentrated solar power. Storage systems support the adaptation of power cycles or industrial processes, permitting usually only slow thermal transients, to the energy flow provided by the solar collectors, which can show very fast variations since only the direct irradiation is used.

Only in the first of the early solar thermal power plants built between 1985 to 1991 in the USA, storage capacity was integrated. The focus in this initial phase was mainly on the development of collector components. Many of the commercial solar thermal power plants being developed or under construction in Spain include storage capacity. In order to minimize financial risk, the most mature storage technologies (molten salt storage or steam accumulators) were chosen for these facilities. In parallel, research activities aiming at increased efficiency, reduced investment costs, and extended operation temperature range have been initiated.

Classification of Thermal Storage Systems

Concentrated solar heat is used for electricity generation and process heat applications. While the maximum temperature of the working fluid strongly depends on the application, the minimum temperature usually exceeds 100°C. Consequently, liquid water at

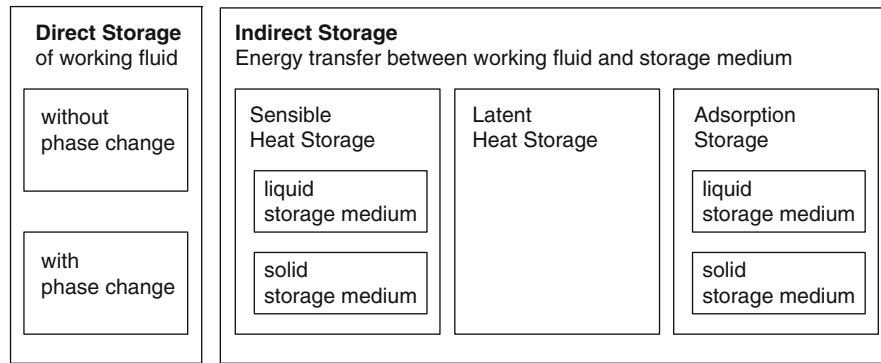
atmospheric pressure cannot be used as storage medium; experiences from low-temperature systems intended for heating and cooling cannot be applied. For medium- and high-temperature thermal energy storage various basic concepts have been suggested. These concepts can be described by various technical criteria. Among these the most important are:

- Thermal power provided during the discharging process, power transferred during the charging process
- Capacity of the storage unit, i.e., total energy provided during the discharge process
- Temperature of heat provided during the discharge, maximum temperature accepted during charging process
- Frequency of charging/discharging cycles
- Reaction time needed to provide nominal load
- Thermal losses, parasitic power needed during charging/discharging process
- Exergetic efficiency

The multitude of criteria makes a direct comparison of storage systems difficult. Values like thermal power or temperature might also be time-dependent for certain storage concepts. Storage units must be adapted both to the energy source (solar collectors) and the consumer (thermal process). Since different types of solar collectors can be combined with various thermal processes, the resulting systems show specific requirements regarding the integration of a storage unit. Consequently, a general rating identification of the most promising concept is not possible; the selection depends on the specific application.

Thermal Energy Storage Concepts

This survey focuses on storage systems operated at temperatures exceeding 100°C and intended for applications requiring thermal power between 100 kW (solar process heat application) and several hundreds of MW (solar thermal electricity). Basic concepts proposed for such applications can be divided into systems applying a direct storage of the working fluid used in the solar collector and indirect systems transferring energy to a separate storage medium as shown in Fig. 6 [21, 22]. The simplest concept uses an additional volume of hot working fluid to store energy. This concept is usually



Thermal Energy Storage. Figure 6

Basic storage concepts under consideration for systems using concentrated solar power

only cost-effective if storage is possible at atmospheric pressure, the partial pressure of the working fluid at storage temperature should be low. In a modification of this concept, the working fluid undergoes a phase change during the charging/discharging process.

Many working fluids cannot be directly stored; the energy must be transferred to a separate storage medium. Dependent on the physical principle used for changing the energy content of the storage material, sensible heat storage can be distinguished from latent heat energy storage and adsorption concepts. While indirect sensible storage has already reached commercial status, latent heat storage has recently reached pre-commercial status. Various latent heat storage materials showing different melting points can be composed to form a cascaded system. Sensible heat storage systems and latent heat storage systems can also be combined to benefit from the specific advantages of both concepts. In addition to the aforementioned concepts, sorption systems using energy changes related to adsorption/desorption processes are considered for energy storage, but for temperatures relevant for concentrated solar power an application seems not to be imminent. This is also true for chemical storage concepts, where questions concerning reversibility and efficiency have to be clarified.

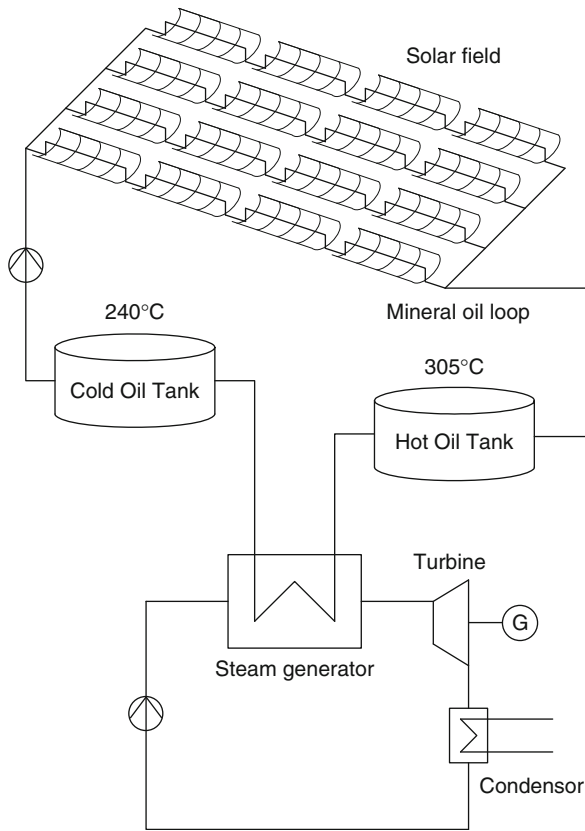
Direct Storage of Liquid Working Fluid

The various storage concepts show different states of maturity. Due to cheap fossil fuel available as

backup source and limited requirements concerning dispatchability, the extent of storage capacity integrated in early commercial solar thermal facilities was limited. For the initial installations, concepts based on a direct storage were preferred representing a low-risk approach demanding limited development effort. The cost efficiency of storage concepts is strongly related to the frequency of charging/discharging cycles. The focus is currently on systems compensating transients caused by clouds and on systems for daily charging/discharging system. Seasonal storage is not considered to be cost-effective within the near future.

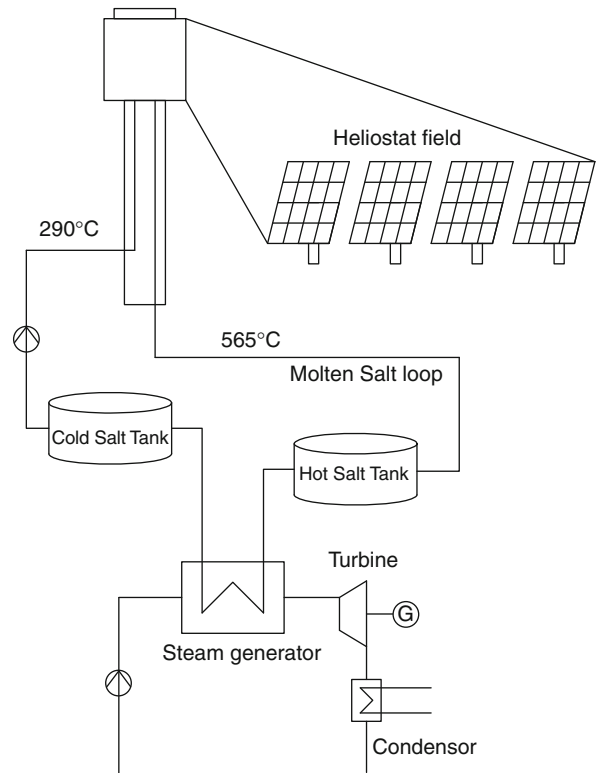
Among the nine parabolic trough power plants SEGS I- SEGS IX built in California between 1985 and 1991, only SEGS-I was equipped with a storage unit allowing the operation of the 14 MW_{el} plant for about 3 h (Fig. 7) [23]. In contrast to the later SEGS-plants, SEGS-I uses a mineral oil as heat transfer medium in the solar collectors, which is significantly cheaper than the synthetic VP1 oil applied in the other SEGS plants. The relatively low-cost mineral oil allowed the direct storage in an additional volume of hot heat transfer fluid. This approach is not cost-effective with VP1 not only due to the higher costs of the heat transfer fluid: At 400°C VP1 requires a pressure of about 11 bar, so a pressure vessel is needed for storage. In SEGS-I about 3,260 m³ mineral oil were stored in the hot tank at a temperature of 307°C. The storage system was damaged by a fire in 1999.

The second example for large-scale direct energy storage is the Solar Two central receiver power plant



Thermal Energy Storage. Figure 7

Simplified scheme of SEGS-I parabolic trough plant with direct storage of heat transfer fluid



Thermal Energy Storage. Figure 8

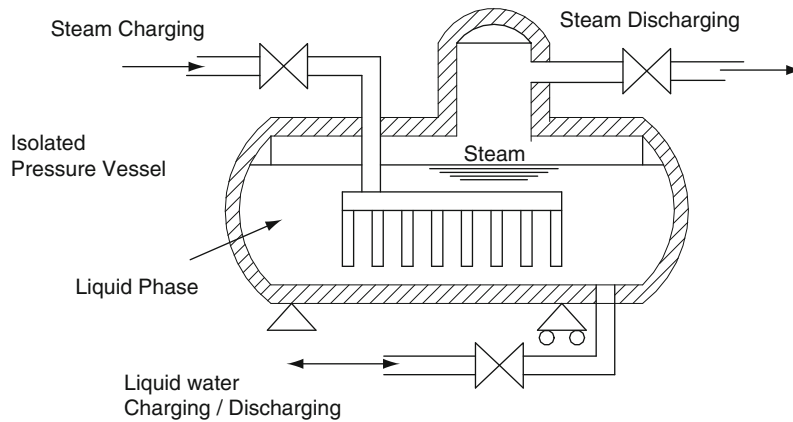
Simplified scheme of Solar Two power plant with central receiver and direct storage of molten salt used as heat transfer fluid

using molten salt as a heat transfer fluid (Fig. 8). This demonstration power plant was erected in 1994 on basis of the Solar One facility and was operated until 1999. The maximum electrical power was 11 MW, the two-tank storage system with a total volume of about 1,700 m³ had an inventory of 1,400 t of a mixture of molten sodium nitrate (60% by weight) and potassium nitrate (40% by weight). The thermal capacity of the storage system operated between 565°C and 290°C was 107 MWh, allowing the operation of the turbine for 3 h [24].

Direct Storage with Phase Change (Steam Accumulators)

While today's application of energy storage in the process industry is still limited, almost the complete

existing capacity is based on steam accumulator technology. Here, the unique thermal storage ability of liquid water is applied by using pressure vessels as storage tanks (Fig. 9). Steam accumulators are charged by condensation of steam fed into the pressurized liquid volume [25, 26]. Since the pressure vessel is filled with saturated liquid water, the pressure increases during the charging process. During the discharge process the pressure in the storage vessel is decreased and saturated steam is extracted. The storage capacity corresponds to the variation of sensible energy of the liquid volume during the discharging process. Since the water in the volume is in the saturation state, a change in temperature is correlated to a change of pressure of the saturated steam provided during the discharge process. While a larger decrease of pressure is advantageous regarding the volumetric storage capacity, the acceptable pressure



Thermal Energy Storage. Figure 9
Scheme of a steam accumulator

variation might be limited by efficiency considerations. Due to the logarithmic correlation between saturation pressure and saturation temperature, the absolute reduction of the pressure during the discharge process is increased at higher pressure levels to provide the same energy. The cost-effective capacity of steam accumulators is limited by the size of pressure vessel. Due to the short reaction time this concept can be used as buffer storage to compensate cloud passing or to support other storage concepts, which show a larger storage capacity and require a longer start-up procedure.

The PS10 central receiver power plant, representing the first European commercial solar thermal power plant starting operation in 2007, uses a steam accumulator for compensation of cloud transients. Four tanks with a total storage capacity of 20 MWh enable a 50% load operation of the 11 MW_{el} for about 50 min. The storage system is charged with saturated steam at 45 bar; during the discharge process saturated steam is provided at variable pressure.

Indirect Storage Concepts

Direct storage is not a cost-effective solution for all working fluids used in solar absorbers. Synthetic heat transfer oil is used to operate the parabolic troughs at temperatures up to 400°C, but this heat transfer fluid is too expensive for direct storage. Superheated steam and hot air show a very low volumetric energy density, so direct storage in pressure vessels is not practical. Storage

systems for these working fluids use a separate storage medium. Pressurized working fluids (synthetic oil, steam) utilize a heat exchanger to transfer the energy between working fluid and storage medium. Efficient indirect energy storage demands the minimization of the temperature difference between the heat transfer fluid and the storage medium. Since both single-phase fluids (e.g., thermal oil, air, molten salt) and two-phase fluids (e.g., steam) are used as heat transfer medium in the solar collectors, the corresponding storage systems use either single-phase storage media (sensible heat storage) or two-phase storage materials (phase change materials, PCM). It is difficult to compare single phase and two phase storage systems on a component level. The storage performance depends on the integration in the overall process. Hence, it is recommended to compare single phase and two phase approaches on a system level. In contrast to low-temperature heating applications, the higher volumetric energy density of PCM systems is usually not relevant as criterion for concentrating solar systems, since necessary space is only of minor importance.

In all indirect storage concepts the maximum temperature of the working fluid is lower during the discharge process than during the charging process due to the unavoidable necessary temperature difference between working fluid and storage material. The minimization of this decrease and cost-effective design of the heat exchanger is a central issue for all indirect storage concepts.

Indirect Storage in Solid Media Sensible heat storage in solid media requires the integration of a heat exchanger into the storage material. While capacity-related costs play an important role in the selection of storage media, other requirements result from the integration of the heat exchanger, since the investment for the tubes and manifolds represents a significant share of total costs. The optimization of the geometry of the heat exchanger is essential for a cost-effective storage system. The thermal heat conductivity and the volumetric storage capacity of the storage material determine the needed surface area of the heat exchanger. Additional aspects are the thermal expansion rate of the storage material, which should be similar to the material used for the heat exchanger to limit thermal stress. The storage material should show a high cycling stability. Considering the various selection criteria, high temperature resistant concrete was identified as promising storage material [1]. Table 5 shows some of the relevant properties.

The storage capacity of sensible storage systems depends on the temperature variation during the charging process: figures for capacity-specific costs for the storage material require the definition of the temperature variation in the storage material during the discharge process. This value is also slightly influenced by the temperature range, since the specific heat capacity is temperature dependent.

During recent years, concrete storage systems intended for parabolic trough power plants using synthetic oil as heat transfer fluid have been developed. This storage unit is operated between 290°C and 390°C with daily charging/discharging cycle (Fig. 10). For this

application, a parallel tube heat exchanger design using a concrete with quartz aggregates and a low level of binder made of blast furnace and flue ash is considered to be the most cost-effective solution. In order to increase the thermal conductivity, several alternative material options have been examined. Additives included metal splinters and graphite powder. So far, these approaches did not show an economic advantage. The integration of heat transfer enhancement structures within the concrete block has also been researched. It was found that the additional costs outweigh the benefit of this approach.

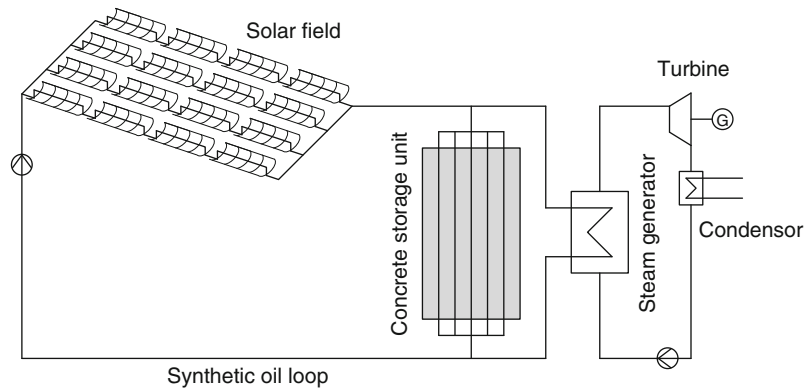
The optimization of the operation strategy offers an option for improving the efficiency of a solid medium storage. If the mass flow rate through the storage and the inlet temperature is kept constant during the charging process, the temperature at the outlet of the storage unit increases dependent on the charging state. Usually, this temperature is limited to avoid damaging of the absorbers. Consequently, the charging must be stopped, the capacity of the storage unit cannot be utilized completely. By applying modular storage units, the usable storage capacity can be increased. An improved adaptation of the storage to the demands of the power cycle during the discharging process results in a further increase of the usable capacity [2].

A test module with a capacity of 350 kWh was connected to parabolic trough collectors of the Plataforma Solar de Almeria and operated successfully during summer 2003. Long-term stability of the concrete storage approach was proven by operation of an enhanced storage module from 2008 to 2010 (Fig. 11). While these test units were operated with synthetic oil as heat transfer medium, concrete storage is also considered for application with superheated steam. A second-generation concrete storage module with a capacity of about 300 kWh is being built and tested for operation with steam at 100 bar and 500°C [27].

Thermal Energy Storage. Table 5 Properties of high-temperature concrete used for sensible heat storage

Density at 370°C (kg/m ³)	2,250
Specific heat capacity at 370°C (J/kgK)	1,100
Thermal conductivity at 370°C (W/mK)	1.3
Coefficient of thermal expansion at 370°C (10 ⁻⁶ /K)	11.6
Capacity (kWh/m ³ ·K)	0.68
Capacity specific costs (\$/kWh) (only storage material, ΔT = 100 K)	1.8

Indirect Storage in Liquid Media Indirect storage in liquid media allows the separation of storage medium and heat exchanger; the size of the heat exchanger is only determined by the necessary power and not by the capacity of the storage unit. Since the area-specific power density is higher, using liquid storage media usually requires smaller heat exchangers than using



Thermal Energy Storage. Figure 10

Simplified scheme of a parabolic trough power plant with integrated concrete storage



Thermal Energy Storage. Figure 11

Photo of a concrete storage module during long-term testing

solid storage media with embedded heat exchanger. A significant advantage of liquid media storage is the simple operation strategy, because the temperature can easily be kept constant during the discharge process.

Table 6 shows the relevant properties for some potential liquid storage media. Regarding costs, safety aspects, and thermal stability within the relevant temperature range, nitrate salts and nitrite salts are the preferred candidate fluids for liquid energy storage. The application of these salts requires the consideration of the lower temperature limit defined by the melting temperature and the maximum temperature limit resulting from thermal stability. Freezing must be

prevented in the piping, the heat exchanger, and in the storage tanks.

The advantage of using molten salt storage systems is the availability of experiences from the Solar Two project. Since this concept is considered as already proven, it was selected for the Andasol power plants using parabolic trough technology [28]. The schematic layout of the plant is shown in Fig. 12.

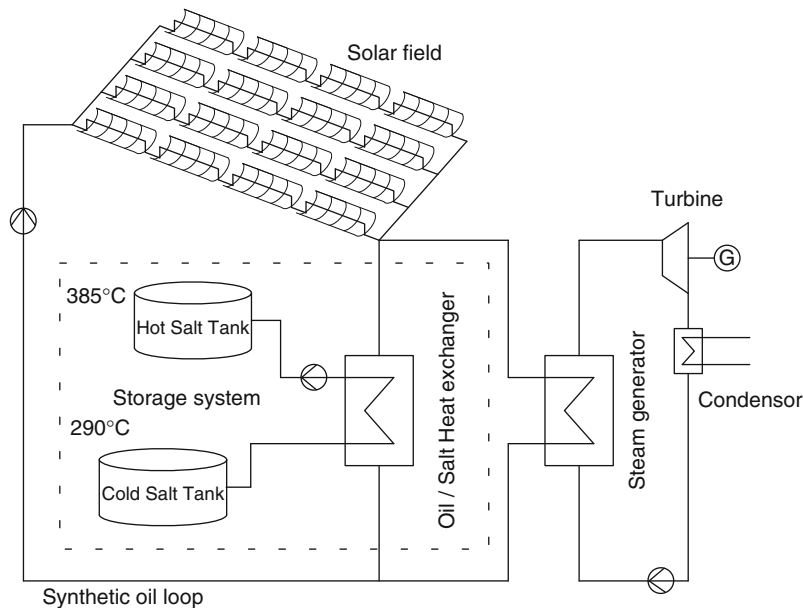
At present, the two-tank molten salt storage is the only commercially available concept for large thermal capacities being suitable for solar thermal power plants. In the Andasol I plant 28,500 t of molten salt (60 wt.% NaNO_3 , 40 wt.% KNO_3) are stored in two tanks with a total volume of 32,600 m^3 and are operated between 385°C and 290°C. It presents a 7.5 h storage system providing energy to generate 50 MW electrical power. Plant operation was started in 2009 [29]. Based on the favorable experience with the Andasol storage systems, the molten salt storage approach is applied to Andasol II and III plants as well to further parabolic trough plants in Spain. Detailed characteristics and experience with the two-tank system are well documented [29, 30].

Limitations of molten salt storage are high costs of the storage media, the high melting point of molten salts that lead to the need for freeze protection and risk of corrosion. Figure 13 shows the main expense fractions. This shows that the salt inventory and the tanks have significant cost-reduction potentials.

However, specific storage costs can be reduced considerably, when the storage is implemented in systems with larger temperature differences, as the

Thermal Energy Storage. Table 6 Data for media considered for liquid storage of sensible heat

	Application range (°C)	Specific heat capacity (J/(kgK))	Density (kg/m ³)	Capacity (kWh/m ³ K)	Specific costs – only storage material (\$/kWh) ($\Delta T = 100$ K)
KNO ₃ –NaNO ₃ (eu)	222–540	1,400	2,000	0.77	12.5
KNO ₃ –NaNO ₂ –NaNO ₃ (eu)	142–540	1,400	2,000	0.77	22.5
Mineral oil	0–320	2,900	700	0.56	42
Synthetic oil (pressurized)	20–400	2,450	750	0.68	58



Thermal Energy Storage. Figure 12

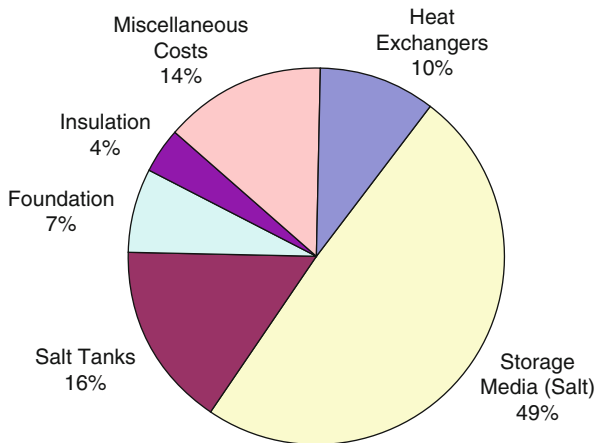
Simplified scheme of a parabolic trough power plant with indirect storage system using molten salt

capacity of a molten salt storage is directly proportional to the temperature difference between the hot and the cold tank. In this context, the commercial solar tower power plant Gemasolar in Andalusia, Spain is an example. The construction of the plant was completed in 2010 and the start of operation is scheduled for 2011 [31]. In this power plant, the first commercial direct two-tank system will be realized.

Indirect Storage with Phase Change Materials

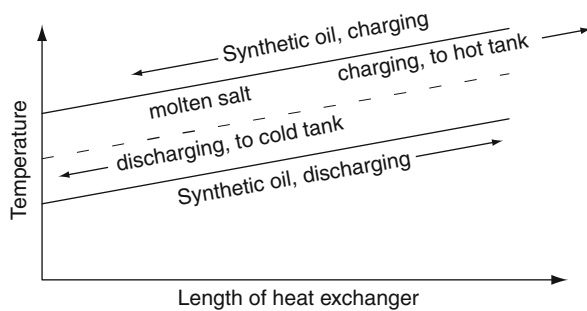
Phase change materials (PCM) are used for energy storage with little temperature variations of the storage material. Most PCM systems use the energy

associated with melting or solidification processes. This energy is transferred at nearly constant temperature, characteristic energy densities of storage materials are in the range of 50–200 kJ/kg. This latent heat corresponds to a change of sensible energy resulting from a temperature change of 100–150°C for a typical storage material. While PCM systems usually show a higher complexity than sensible heat storage units, they are preferred for systems using steam in absorbers, since during the isothermal condensation and evaporation processes the temperature difference between steam and storage material can be kept small (Figs. 14, 15).



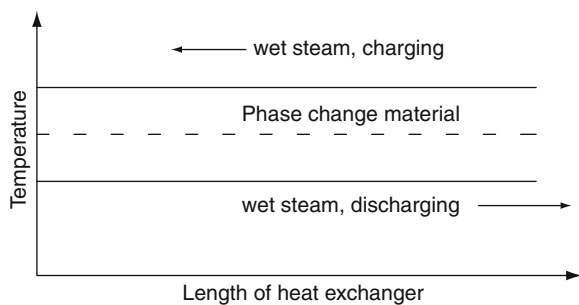
Thermal Energy Storage. Figure 13

Distribution of investment costs for an indirect two-tank storage system [28]



Thermal Energy Storage. Figure 14

Heat transfer in an indirect storage system using synthetic oil as heat transfer fluid and molten salt as storage medium



Thermal Energy Storage. Figure 15

Heat transfer in an indirect storage system using saturated steam as heat transfer fluid and PCM as storage medium

Thermal Energy Storage. Table 7 Selected materials with melting temperatures in the range relevant for solar thermal power plants and solar process heat applications using steam

Storage media	Melting temperature (°C)	Capacity-specific costs (\$/kWh) only material, only latent heat	Heat conductivity (W/[m K])
Tin	232	900	60
Lead	327	150	35
Nitrate/nitrite salts	130–350	20	0.5

A first selection of candidate storage materials results from the melting temperature, which should be adapted to the specific process. In systems using steam as working fluid, the melting temperature should be close to the saturation temperature resulting from the operation pressure. For concentrating solar systems with operation pressures in the range 1–120 bar, the required melting temperatures are between 100°C and 325°C. Tables 3 and 7 show selected materials with melting temperatures in this range.

With respect to costs, nitrate and nitrite salts and their mixtures are considered to be the most promising candidate materials for PCM storage. These salts can be applied either as pure substances or as (eutectic) mixture of two or more components. While a large number of systems have been suggested as PCM [32], due to aspects like environmental compatibility, hygroscopicity, or cycling stability, a limited number of substances are considered to be of practical use; three examples are shown in Table 8.

The relatively poor thermal conductance represents the major drawback of these salts. This is the dominant problem complicating the implementation of cost-effective PCM systems due to low volumetric power density. PCM systems exceeding temperatures of 100°C have long been limited to lab-scale experiments. If a heat exchanger is embedded in the storage volume, pure PCM requires a very large heat transfer area resulting in high investment costs. The successful demonstration of direct steam generation

Thermal Energy Storage. Table 8 Properties of selected nitrate salts as PCM

Storage media	Melting temperature (°C)	Heat of fusion (kJ/kg)	Density (kg/m ³)	Capacity latent heat (kWh/m ³)	Costs (\$/kWh) only material, only latent heat
KNO ₃ -NaNO ₃ (eu)	222	100	2,000	55	17.5
KNO ₃ -NaNO ₂ -NaNO ₃ (eu)	142	60	2,000	33	52.5
NaNO ₃	306	175	1,900	96	9.5

in parabolic troughs gave rise to an increased demand for PCM storage. The focus of the research activities during the last years [13] were directed to:

- Macro-encapsulation of PCM in small containers stored in pressurized steam volume
- Composite materials made of PCM and an additional material showing high thermal conductivity (e.g., graphite)
- Integration of heat transfer structures into the storage volume to enhance the heat transfer within the PCM

The integration of an internal heat transfer structure has been identified as the most promising concept. The so-called Sandwich concept uses thin plates arranged perpendicular to the axis of tubes of a parallel pipe heat exchanger embedded in the PCM volume. Candidate plate materials are expanded graphite showing a high thermal conductivity (>150 W/(mK)) or metals.

A pilot storage for solar plants with direct steam generation was designed and constructed using sodium nitrate (NaNO₃) with a melting temperature of 306°C as the phase change material. The PCM storage module uses the sandwich design with aluminum fins to reach the required high heat transfer rates and accordingly the required charge and discharge times. After successful testing of the lab-scale storage, the design was scaled up to about 14 t of salt with a latent heat capacity of approx. 700 kWh and adapted to the design parameters of the water/steam test loop (128 bar, 400°C) [33]. A picture of the PCM module is shown in Fig. 16. In contrast to previous test modules, the tube register is arranged vertically. As the sodium nitrate experiences a volume increase of approx. 10% during melting, an excess volume is provided for the liquid phase at

**Thermal Energy Storage. Figure 16**

Photo of a 700 kWh NaNO₃ PCM pilot storage with internal recirculation equipment

the top of the storage. It needs to be ensured that there are open paths for the fluid to expand into this volume, which is accomplished by the vertical position of the tubes, where the salt melts first. Pressure losses

through the PCM storage module are very low, as the fluid flows through many parallel pipes of ca. 7 m length, resulting in low flow velocities.

In charging mode, steam with a temperature slightly above saturation properties (typically ~ 107 bar, $\sim 320^\circ\text{C}$) is routed into the PCM module where it condenses. The flow direction during charging is from top to bottom so that the condensate is removed by gravity. A condensate drain assures that the medium leaves the module only in liquid form. The module is expected to be able to condense the full mass flow of 0.8 kg/s that the test loop can provide.

For discharging, the tube register of the PCM module is flooded with liquid water at a temperature just below saturation properties (typically ~ 81 bar, $\sim 295^\circ\text{C}$). The saturated steam produced by the module during discharge is dried in a steam separator (spherical drum on the tower to the left of the insulated PCM module in Fig. 16). The liquid water from the steam separator is recirculated either by natural recirculation or by a pump. During discharge, the flow direction is from bottom to top.

If superheated steam is generated in the solar collectors, both single phase (preheating, superheating) and two-phase (evaporation) heat transfer takes place. The corresponding storage system must allow both sensible and isothermal storage to be efficient. The

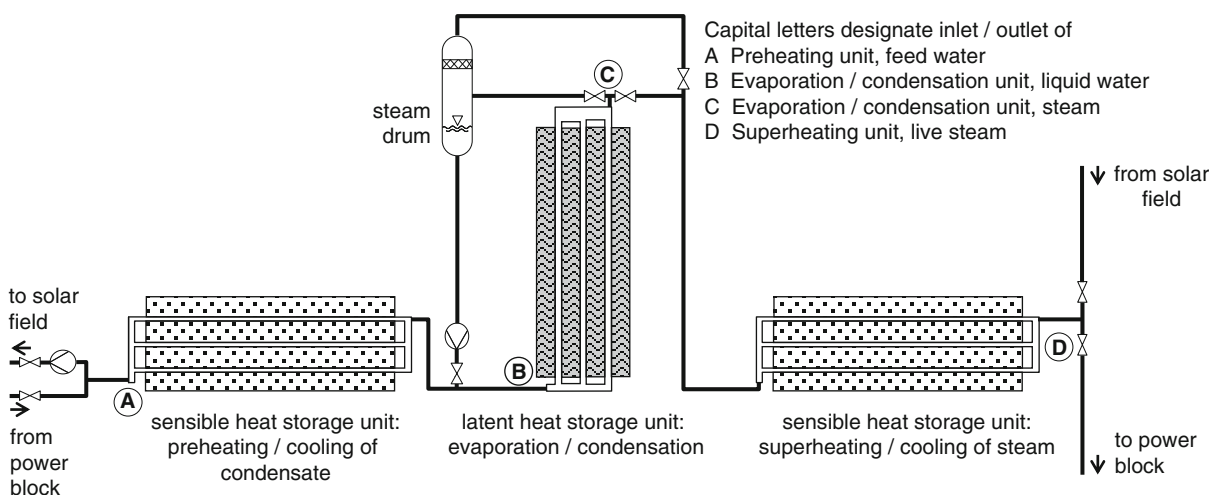
conceptual layout of such modular storage system for storage and generation of superheated steam is shown in Fig. 17.

Regenerator Storage for Solar Central Receiver Plants

Central receiver power plants using air as a heat transfer medium offer the potential to particularly cost-effective solutions with high conversion efficiencies [34–36]. Also, the suitable heat storage technology – regenerator storage based on directly heated solid media – has a simple setup, is applicable to highest temperatures and has best prospects for a deployment in large installations [37, 38]. These aspects indicate good opportunities for a near-term commercialization.

However, for a successful market introduction, efficient and up-scalable solutions for the heat storage are a prerequisite. Storage regenerators are well suited to the needs, but need further investigations to clarify open questions concerning design and materials. Though first concepts have been looked at in the early 1990s, the status of this storage application is still at an early stage.

Initial designs and tests for atmospheric receivers have been carried out by the German TSA consortium.



Thermal Energy Storage. Figure 17

Overview of a modular thermal energy storage system for DSG combining sensible and latent heat storage

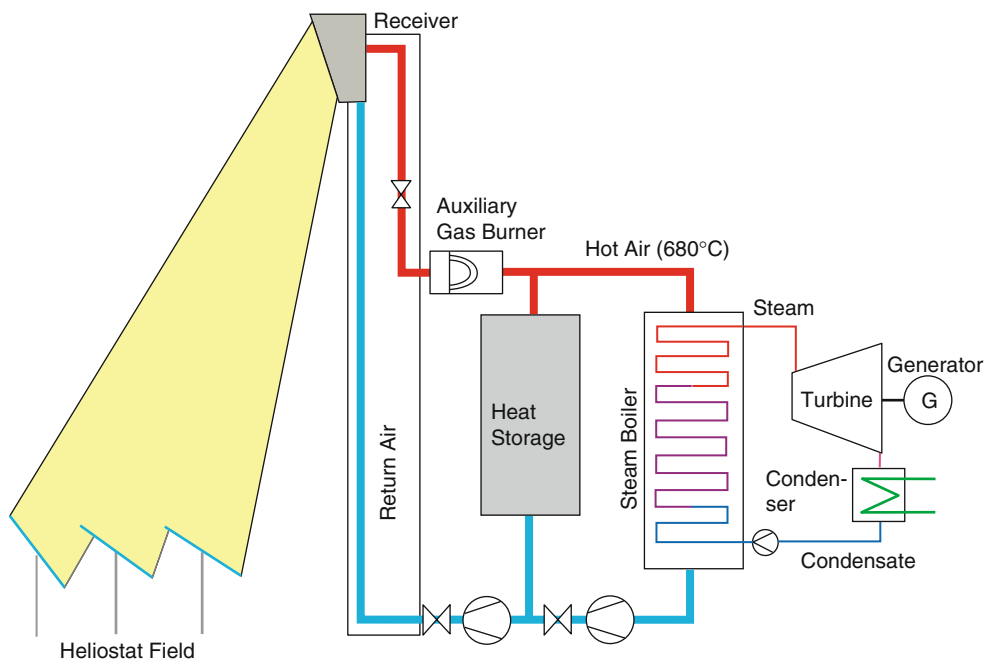
For that purpose, a test facility, including a thermal storage, was erected in Spain in 1991. However, rather than aspects of storage design, the system demonstration was in the focus of work. The storage was based on a packed bed of aluminum oxide in spherical shape, providing a thermal capacity of about 1 MWh. Though fully functional, this concept turns out not to allow a direct scale-up to commercial size: high costs for the high-purity aluminum oxide and increased thermally induced mechanical loads during cyclic operation represent the main barriers.

The heat storage of the recently commissioned solar tower plant in Jülich, Germany uses refractory in a stacked arrangement and therefore avoids such mechanical issues [39]. The Solar Power Tower Jülich is an experimental central receiver plant, inaugurated in 2009 to serve the needs of a further development of this technology. Erection and operation of the plant are accompanied by a research program with the objective to cover remaining uncertainties of design and operation and to further promote the development of the technology toward

a commercial deployment [39, 40]. As a part of this work program also heat storage operation and design are addressed.

The plant uses an open volumetric receiver technology. In its primary cycle, air at atmospheric pressure is heated up to temperatures of about 700°C. This solar heat then powers a steam generator, producing steam at 100 bars and 500°C and driving a 1.5 MW_{el} turbine-generator set. In parallel to the steam generator and receiver, the heat storage is integrated into the power cycle, see Fig. 18 below. It is implemented as an air-cooled regenerator storage, an installation that is still unique in this application.

The design of the storage was derived from industrial applications of thermal exhaust air purification. These RTO (regenerative thermal oxidizer) systems are applied to purify air from volatile organic compounds through an exothermic process taking place inside a regenerator's heat exchanging ceramic inventory at temperatures up to 1,100°C. The storage is located within the plant's tower. It extends over two stories, close to the locations of receiver and steam generator.



Thermal Energy Storage. Figure 18

Schematic plant layout of solar power tower with integrated regenerator storage

At rated operation conditions, the storage system is cycled between 120°C and 680°C and supplies a storage capacity of almost 9 MWh. Further technical specifications include data for the discharge heat rate, discharge duration, and permissible outlet temperature variations as summarized in Table 9.

During charge operation the hot air from the receiver enters the storage's dome and flows downward through the storage chambers, forming a moving

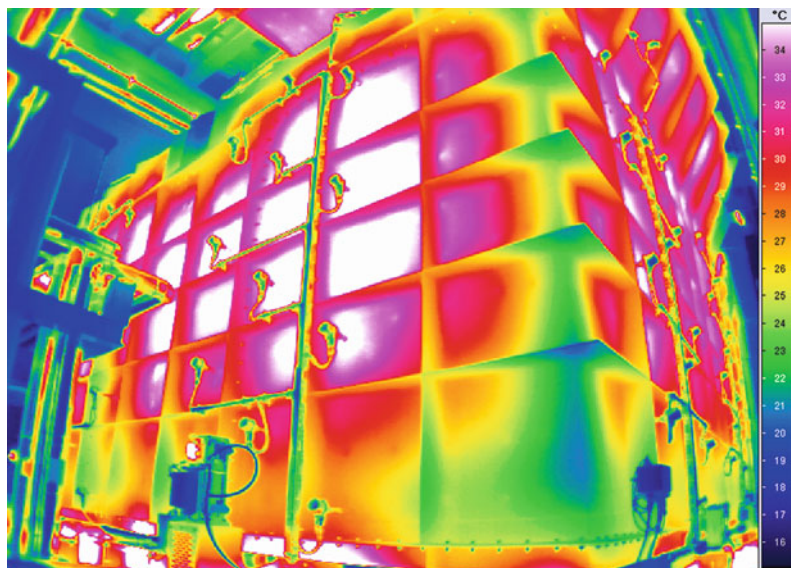
temperature profile as typically found in regenerator storage. Reversing the flow direction through the storage starts discharge operation and supplies heat to the steam generator's air loop.

Full functionality of the storage subsystem was confirmed during a test campaign performed in 2010: Standstill operation shows that thermal losses are kept at a low level (Fig. 19). Cycling operation at high heat rates levels could be performed as expected for charge and discharge operation. The successful installation and testing experience is considered an important step toward a further development of this storage technology. Present results support the view that the technology represents a promising basis for up-scaled implementations in solar applications and beyond.

On that way, further aspects need to be addressed. This includes a better understanding of internal air flow and flow distribution, design improvements in terms of heat transfer, losses, and cost efficiency, as well as the investigation of alternatives for the inventory arrangement, including packed beds and the potential for higher operation temperatures [40, 41].

Thermal Energy Storage. Table 9 Specifications of an air-cooled regenerator storage

Inlet temperature (charge/discharge)	680°C/120°C
Outlet temperature (charge/discharge)	680–640°C/120–150°C
Charge mass flow	9.4 kg/s
Discharge heat rate	5.7 MW _{th}
Full load discharge period	1.5 h
Pressure loss	<1,500 Pa



Thermal Energy Storage. Figure 19

Infrared image of the lateral storage surface during standstill

Future Directions

Current research topics are aiming at the development of advanced thermal energy storage systems having the capability to reduce the investment cost. New approaches are covering heat transfer fluids and storage materials with superior thermophysical properties as well as new concepts for improved heat transfer and storage design. Examples are:

- Binary and ternary liquid salts systems with low freezing point but thermal stability beyond 600°C
- Low-cost filler materials to replace expensive salt systems
- Thermally stable nanomaterials to improve thermophysical properties of liquids
- PCM materials with thermal conductivity beyond 10 W/mK
- New heat transfer concepts to decouple the functions of storage capacity and loading power
- Sensible heat storage with fluidized solid particles
- Adjustment of thermochemical heat storage for high-temperature application

Most of these approaches are still in the laboratory scale and need to be validated and proven for up-scaling to commercial scale. From the results important impact toward more efficient, reliable, and economic thermal energy storage systems can be expected.

Bibliography

1. Laing D, Steinmann W-D, Tamme R, Richter C (2006) Solid media thermal storage for parabolic trough power plants. *Sol Energy* 80:1283–1289
2. Laing D, Steinmann W-D, Fiß M, Tamme R, Brand T, Bahl C (2008) Solid media thermal storage development and analysis of modular storage operation concepts for parabolic trough power plants. *J Sol Energy Eng* 130:011006-1/5
3. Goldstein M (1961) Some physical chemical aspects of heat storage. In: U.N. Conference on new sources of energy, Rome, vol 35, pp 5–7
4. Telkes M (1974) Solar energy storage. *ASHRAE J* 16:38–44
5. Altmann M, Yeh H, Lorsch HG (1973) Conservation and better utilization of electric power by means of thermal energy storage and solar heating. Final summary report, NSF/RANN/SE/G/27976/PR 73/5, University of Pennsylvania
6. Lorsch HG (1974) Thermal energy storage. Final report, NSF/RANN/74-021C
7. Carlson B, Stymme H, Wettermark G (1978) Storage of low-temperature heat in salt-hydrate melts – calcium chloride hexahydrate. Swedish Council for Building. Research D 12
8. Ozawa T et al (1980) Screening of latent heat thermal energy storage materials by using evaluated thermodynamic data. In: 7th Codata international conference, Kyoto
9. Mar RW (1980) Material science issues encountered during the development of thermochemical concepts. In: Murr LE (ed) *Solar materials science*. Academic, London
10. Mehling H, Cabeza LF (2008) Heat and cold storage with PCM – an up to date introduction into basics and applications. Springer, Heidelberg
11. Bauer T, Laing D, Steinmann WD, Kröner U, Tamme R (2008) Screening of phase change materials for process heat applications in the temperature range 120 to 250°C. In: *Proceedings of Eurosun, Lisbon, 7–10 Oct 2008*
12. Tamme R, Bauer T, Buschle J, Laing D, Müller-Steinhagen H, Steinmann W-D (2008) Latent heat storage above 1200 C for applications in the industrial process heat sector and solar power generation. *Int J Energy Res* 32:264–271
13. Steinmann W-D, Tamme R (2008) Latent heat storage for solar steam systems. *J Sol Energy Eng* 130:011004-1/5
14. Wentworth WE, Chen F (1976) Simple thermal decomposition reactions for storage of solar thermal energy. *Sol Energy* 18:205–214
15. Schaub F, Wörner A, Tamme R (2010) High temperature thermo-chemical heat storage for CSP using gas-solid reactions, *Proceedings of SolarPACES 2010, Perpignan*
16. Wong B, Brown L, Schaub F, Tamme R, Sattler C (2010) Oxide based thermochemical heat storage, *Proceedings of SolarPACES 2010, Perpignan*
17. Stobbe ER, de Boer BA, Geus JW (1999) The reduction and oxidation behaviour of manganese oxides. *Catal Today* 47:161–167
18. Zaki MI et al (1998) Thermochemistry of manganese oxides in reactive gas atmospheres: Probing catalytic MnOx compositions in the atmosphere of CO+O₂. *Thermochimica Acta* 311: 97–103
19. Lovegrove K et al (2004) Developing ammonia based thermochemical energy storage for dish power plants. *Sol Energy* 76:331–337
20. Buck R et al (1994) Development of a volumetric receiver-reactor for solar methane reforming. *J Sol Energy Eng* 116: 73–78
21. Beckmann G, Gilli PV (1984) *Thermal energy storage*. Springer, Berlin
22. Dinter F, Geyer M, Tamme R (1990) *Thermal energy storage for commercial applications*. Springer, Berlin
23. Herrmann U, Kearney D (2002) Survey of thermal energy storage for parabolic trough power plants. *J Sol Energy Eng* 124:145–152
24. Pacheco JE (2002) Final test and evaluation results from the solar two project. Sandia National Laboratories, SAND2002-0120

25. Goldstern W (1970) Steam storage installation. Pergamon Press, Oxford
26. Steinmann WD, Eck M (2006) Buffer storage for direct steam generation. *Sol Energy* 80:1277–1282, Elsevier
27. Laing D, Bahl C, Fiß M (2010) Commissioning of a thermal energy storage system for direct steam generation. In: *Proceedings. SolarPACES 2010*, Perpignan, 21–24 Sep 2010
28. Kelly B, Kearney D (2006) Thermal storage commercial plant design study for a 2-tank indirect molten salt system. NREL/SR-550-40166
29. Relloso S, Delgado E (2009) Experience with molten salt thermal storage in a commercial parabolic trough plant. In: *Proceedings of the SolarPACES 2009*, Berlin
30. Kolb GJ (2010) Evaluation of annual performance of 2-tank and thermocline thermal storage system for trough plants. In: *Proceedings of the SolarPACES 2010*, Perpignan
31. Online newspaper library of Torresol Energy (2011) The innovative Gemasolar plant will be the main project of Torresol Energy at WFES 2011. Press release 11.01.2011. <http://www.torresolenergy.com/TORRESOL/Press/torresol-takes-part-wfes-2011>. Accessed 14 Mar 2011
32. Janz GJ et al (1979) Physical properties data compilation relevant to energy storage II. Molten salts: data on single and multi-component salt systems. NSRDS-National Standard Reference Data System
33. Laing D, Bahl C, Bauer T, Lehmann D, Steinmann W-D (2010) Thermal energy storage for direct steam generation. *Sol Energy* 85:627–633
34. Romero M et al (2002) An update on solar central receiver systems, projects, and technologies. *ASME J Sol Energ Eng* 124:98–108
35. Pitz-Paal R et al (2005) European concentrated solar thermal road-mapping (ECOSTAR): roadmap document. SES-CT-2003-502578
36. Price H (2002) Assessment of parabolic trough and power tower solar technology cost and performance forecasts. Sargent & Lundy, NREL/SR-550-34440
37. Haeger M et al (1994) Phoebus technology program solar air receiver (TSA). Operational experiences with the experimental set-up of a 2.5 MWth volumetric air receiver (TSA) at the plataforma solar de Almería. PSA-TR02/94
38. Fricker HW (2004) Regenerative thermal storage in atmospheric air system solar power plants. *Energy* 29:871–881
39. Zunft S, Hänel M, Krüger M, Dreißigacker V, Göhring F, Wahl E (2010) Jülich solar power tower – Experimental evaluation of the storage subsystem and performance calculations. In: *Proceedings of the SolarPACES 2010 conference (SolarPACES 2010)*, Perpignan, 21–24 Sep 2010
40. Zunft S, Hänel M, Krüger M, Dreißigacker V (2009) High-temperature heat storage for air-cooled solar central receiver plants: a design study. In: *Proceedings of the SolarPACES 2009*, Berlin, 15–18 Sep 2009
41. Dreißigacker V, Müller-Steinhagen H, Zunft S (2010) Thermo-mechanical analysis of packed beds for large-scale storage of high temperature heat. *Heat Mass Transf* 46:1199–1207

Thermal Treatment of Waste: Key Element for Sustainable Waste Management

PAUL H. BRUNNER

Institute for Water Quality, Resource and Waste Management, Vienna University of Technology, Vienna, Austria

Article Outline

Glossary

Definition of the Subject and Its Importance

Introduction

Today's Material Turnover Results in Large Volumes of Waste Requiring Appropriate "Final Sinks"

Goals of Waste Management and of Thermal Waste Treatment

Thermal Waste Treatment for Environmental Protection

The Contribution of Thermal Waste Treatment to Resource Conservation

Future Directions

Bibliography

Glossary

Atmophilic Characterizing an element that concentrates in the atmosphere.

Dioxin Short for di-benzo-dioxine, often substituted with various chlorine atoms such as 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD).

Di-benzo-furan Dioxin with one of the two oxygen atoms replaced by a carbon–carbon bond.

PVC Polyvinylchloride

Transfer coefficient k_{xi} Coefficient describing the partitioning of an element X among "i" products of incineration.

SEA Statistical entropy analysis (see Waste-to Energy (WTE): Decreasing the Entropy of Solid Wastes and Increasing Metal Recovery in this volume).

MSW Municipal solid waste, representing mixed waste that is collected by a given collection system. Since, in addition, other collection systems for recycling might be present too (paper, glass, metals etc.), MSW from different regions with different collection systems might vary even if consumption patterns are identical.

Definition of the Subject and Its Importance

Thermal treatment of waste is not an isolated process: It is part of waste management, energy supply, resource management, and environmental protection. It is linked to economic activities and requires financial, material, and human resources. Without thermal waste treatment, waste management cannot reach its goals. In fact, waste-to-energy (or incineration as it is called commonly in Europe) reduces significantly environmental pollution by persistent organic substances and, also, by some inorganic elements such as heavy metals. Hence, it is important to point out the contribution of waste incineration to sustainable waste management, and to show the potential and limitations of this technology in a broader view. This is even more relevant since the material turnover of modern societies has increased dramatically during the last century, making it necessary to apply powerful and safe technologies both at the front and back end of the material system.

Introduction

Waste management is an integral part of materials management. It serves as an important interface between the so-called anthroposphere (the man-made sphere or technosphere) and the environment [1]. Waste management ensures that harmful substances do not cross this border and that useful materials and energy are cycled back to the anthroposphere. In order to take effective and efficient measures in waste management, a certain knowledge of the material metabolism of the anthroposphere is necessary. A striking example is forecasting of waste generation: When the buildings stock and age distribution, plus certain economic parameters of a region, are known, the generation of construction wastes can be predicted. As another example, when the consumption, stocks, and lifetime of plastic materials are known, the amount and composition of plastic wastes to be treated by recycling and waste disposal can be easily assessed.

Also, in the case of hazardous substances, it is important to have a minimum knowledge of the anthroposphere: Which are the significant sources of heavy metals such as cadmium, lead, and mercury in wastes, and when and how legislation – such as a ban on mercury in batteries – can effect waste composition?

Due to progress in science and technology, the anthropogenic metabolism is constantly changing, with new materials being introduced and increasing complexity of everyday goods. It is therefore necessary to look at the anthroposphere as a dynamic organism with rapidly varying flows (years) and slowly varying stocks (decades). This is also reflected in the amounts and composition of the generated wastes.

Hence, this entry starts with an overview of the material flows and stocks of the most important activities of the metabolism of the anthroposphere, emphasizing changes over time. It then focuses on the goals of waste management and points out the contribution of thermal treatment to waste management. While the focus is on municipal solid waste (MSW) and MSW combustion with energy recovery, most insights can be applied to the thermal treatment of other types of wastes. In particular, the goals of waste incineration, which are explicitly discussed in section “[Goals of Waste Management and of Thermal Waste Treatment](#)”, are the same for all wastes.

The main objective of this entry is to point out that waste incineration is an indispensable constituent of every modern waste management system, because of the variety and complexity of materials used in the anthroposphere: While certain fractions can be recycled, there still remain large amounts of useless and hazardous materials that have to be converted into minerals by means of thermal treatment of the MSW.

Today's Material Turnover Results in Large Volumes of Waste Requiring Appropriate “Final Sinks”

The activity “to clean,” defined as the separation of the “wanted” from the “unwanted,” is a basic human need [1]. All organisms have to take up and excrete materials as a consequence of their metabolism. The food, drinks, and air consumed annually by human beings is about 6 Mg per capita; the resulting output amounts to about 1 Mg per capita of feces and urine and 5 Mg per capita of gas. This so-called physiological metabolism has been basically the same over thousands of years. Even with dietary habits evolving over time and representing differences in culture and economic development, the basic human mass flow is comparatively constant. However, the composition of this mass flow can change substantially, e.g., as

has been observed for the consumption of animal protein in affluent countries.

In addition to the physiological metabolism, modern man depends on the so-called anthropogenic or man-made metabolism. This includes the total material turnover of all goods necessary to sustain a modern standard of living. With the evolution from hunters and gatherers to today's service-oriented affluent societies, the material turnover of men has increased dramatically. Table 1 presents the per capita flows associated with the essential activities of humanity, in prehistoric and modern times. For the activities "to nourish," "to clean," and "to reside," only flows and stocks within private households are shown; the activity "to transport and communicate" also includes the fraction of materials used outside the household, such as roads and other infrastructure.

"To nourish" includes storage, preparation, and consumption of food, beverages, and air. "To clean" stands for the separation of wanted and unwanted goods in households and includes all cleaning processes such as laundry, dish washing, bathroom, vacuum cleaning, etc. The main material used is water mostly as a conveyor belt for dirt, feces, and urine. Materials included in "to reside and work" are mainly fuel (oil, gas, coal, and wood) and construction materials used to protect and shelter humans from the environment. The same materials are instrumental for the activity "to transport and communicate," comprising the transport of persons, goods, information, and energy by individual and public transport media on roads, rails, water, and airways, as well as cables, computers, and satellites.

While the total material turnover has increased by more than one order of magnitude, there has been little change in the "to nourish" activity while the activity "to clean" has experienced an extraordinary increase in mass flows, from less than 0.1 Mg to 60 Mg per capita per year. The same enormous increase has been experienced in the activities "to reside and work" and "to transport and communicate." In modern societies, the activity "to clean" implies the biggest material turnover. Due to the consumption of 60 Mg water per capita annually, modern households and cities experience a healthy environment, with no threats any more from waterborne diseases. The other side of the coin is the large amount of water and wastewater that have to be managed. It becomes evident that for the functioning of modern cities, wastewater management of 60 Mg per capita is more demanding than solid waste management of "only" 2.7 Mg per capita per year (Table 2).

Table 2 below shows the large material stocks that have been accumulated in order to fulfill human requirements. The growth of these stocks is impressive: If present growth rates continue in the future, the anthropogenic stock will double in about 100 years. However, it is doubtful that such a future growth is a realistic assumption. It has to be kept in mind that these activities serve human needs. While a certain material turnover supports human welfare and prosperity, too high a stock may become a heavy burden in terms of resource and financial requirements for maintenance, renewal, and upgrading. It is not known yet if there is an optimum size of anthropogenic stock beyond which the societal benefit becomes marginal or even negative.

Thermal Treatment of Waste: Key Element for Sustainable Waste Management. Table 1 Turnover and stock of materials for prehistoric and modern men (Data from [1])

Activity	Material flows (Mg/capita/year)		Material stocks (Mg/capita/year)		Annual change in material stock (Mg/capita/year)	
	Prehistoric	Modern	Prehistoric	Modern	Prehistoric	Modern
"To nourish"	5.6	5.7	0	<0.1	0	+ <0.1
"To clean"	<0.1	60	0	0.1	0	+ <0.1
"To reside and work"	<0.1	10	<0.1	100	0	+ 1
"To transport and communicate"	0	10	0	160	0	+ 2
Total	6	86	<0.1	260	~0	+ 3

Thermal Treatment of Waste: Key Element for Sustainable Waste Management. Table 2 Input, output, stocks, and stock changes of the most important activities of humanity, in Mg per capita per year [1]

Activity	Input	Output			Stock
		Wastewater	Off-gas	Solid residues	
"To nourish"	5.7	0.9	4.7	0.1	<0.1
"To clean"	60	60	0	0.02	0.1
"To reside and work"	10	0	7.6	1	100 + 1
"To transport and communicate"	10	0	6	1.6	160 + 2
Total	86	61	19	2.7	260 + 3

The data in Table 2 also shows implications relevant for waste management: The solid output of human activities in private households amounts to 2.7 Mg per capita annually. Another 3 Mg per capita are accumulated in anthropogenic stocks. In the future, when these 3 Mg per capita per year have reached the end of their life, they will become wastes. Therefore, it is clear that the amount of wastes to be managed will increase considerably in the future. This is particularly the case for wastes from construction and infrastructure. New measures to reduce the generation of wastes will not much influence the existing material stocks that inevitably will become wastes to be managed.

Table 2 shows the mass flows of materials used in various human activities. However, the quality of the materials used must also be considered. Today's consumer products and long lifetime goods (infrastructure, buildings) are highly complex mixtures of innumerable individual substances. These material combinations allow new and better functions of products, prolong the lifetime of goods, and improve their economic and environmental performance. Since 1939, the world production of chemicals has increased 400-fold, reaching close to 500 million Mg in 2010. As of this date, there are about 7 million known chemical substances. Close to 70,000 are produced for regular use in the industrial, agriculture, and service sector and of these, 5–10% are considered hazardous and about 200 are known to be carcinogenic.

Historically, the materials used were mainly of geogenic (stone, gravel, sand) and biogenic (wood, cotton, and other biomass) origin. Today, while stones, gravel, and sand are still used widely, most products contain metals and synthetic, fossil-based organic

substances. As a consequence, during the last three centuries, wastes produced have changed from harmless geogenic and biomass materials to complex blends of degradable as well as persistent, man-made organic and inorganic materials. While in preindustrial times, hygienic problems associated with waste and waste water were of primary relevance, today's challenge has shifted: The biggest threat for men and environment from modern material turnover and resulting wastes are risks emanating from inappropriate handling of hazardous substances.

Thus, for modern anthropogenic metabolism, waste management must play a key role in

- Redirecting the huge amount of materials used by modern men back into a useful cycle (recycling)
- Providing facilities to safely dispose of the hazardous substances that are leaving the consumption process (safe disposal)
- Informing the manufacturing process regarding the use of substances and design of products (design for recycling and disposal)

In particular – since all materials extracted from the earth and the biosphere, or synthesized by humans, will turn into waste at one point – waste management must provide final sinks for these materials [2]. A "sink" stands for a process that receives waste materials and emissions from human habitat. The most important and powerful sink is the atmosphere which is a perfect conveyor belt for gaseous substances. Next comes the hydrosphere that collects salts and organic substances by rainwater in watersheds and transfers them into rivers that convey them to the sea. The soil also acts also as a sink for landfilled wastes, when sewage sludge

is applied to soil, and for airborne dust particles that settle on land.

Substances may be transported from one sink to another. Over long time periods, speciation of substances in soils, lakes, sediments, or landfills can change significantly due to natural as well as anthropogenic impacts, e.g., by means of eutrophication, acidification, or a shift in redox conditions. In addition to biogeochemical reactions, geological processes such as erosion, weathering, seismic activities, and landslides can relocate substances from one sink to another. Therefore, a sink such as a landfill may not be a long-term sink from a geological perspective.

Sustainable waste management requires that waste-derived problems are solved here and now, and that they are not deferred and passed on to the next generations. Thus, a so-called final sink is defined as a place on the planet where substances have residence times of over 10,000 years. This ensures that the resulting emissions are small and that the fate of a substance is more or less under control. Final sinks are, for instance, underground salt mines, ocean sediments, places on the ground where sedimentation prevails over erosion, etc.

For organic substances, there is a second category of final sinks. In contrast to passive natural final sinks, active “final sink” reactions transform organic substances rapidly into other substances, such as metabolites. In the case of mineralization, the end products are carbon dioxide, water, and –depending on the composition of the initial material – other products such as hydrochloric acid or sulfates. From the point of view of organic substances, such active final sinks represent the last step in the lifetime of a substance. Mineralization is the ultimate method to dispose of hazardous organic substances; it is thus a necessary element of every waste management system.

For affluent societies with a high material turnover, it is necessary to have appropriate final sinks for all substances. The search for ultimate sinks has a long tradition and has not reached its goal as yet [3]. The need for final sinks is well documented by current challenges such as climate change, or the ozone hole, both of which are manifestations of inadequate final sinks. The utilization of fossil carbon overloads carbon sinks, resulting in the rise of concentrations of carbon dioxide and methane in the atmosphere. Also, production, use, and emissions of certain chlorinated and

fluorinated carbohydrates (CFCs) have resulted in high concentrations of CFCs in the sink stratosphere, depleting the ozone layer and creating the so-called ozone hole.

Goals of Waste Management and of Thermal Waste Treatment

The first goal of waste management has been and still is the *protection of human health and life*. While in affluent countries with well-developed waste management systems this has become a matter of course, in many less developed countries this is still the chief objective for establishing waste management. In parallel to this goal, and also linked to it since human health depends on a sound environment, stands the *protection of the environment*.

The second goal of waste management is *resource conservation*: conservation of materials, energy, and land. This objective becomes the more important with higher material turnover and population density, with megacities as hotspots of resource use and resulting wastes. There are two main benefits from the pursuit of this goal: Recycling enhances the availability of resources and significantly reduces environmental pollution by reducing the need for primary production that in general represents the most polluting step in the life cycle of a product.

The third goal represents more of a framework for sustainable waste management: The practice of waste management shall *not cause any burden for future generations*. Landfilling that produces long-term emissions requiring treatment and control for several centuries [4] does not comply with this objective. What is needed are final storage landfills that have very small emissions even for long time periods (thousands of years). Landfills can only become final sinks for substances when wastes are pretreated and organic compounds are completely mineralized, before landfilling. Recycling of wastes containing hazardous substances without removing the harmful constituents such as heavily stabilized plastic materials also poses a burden for future generations.

Concerning neighboring countries, wastes should be exported *only* when the receiving country is able to manage wastes as well as its constituents in a safe, responsible, and resource-conserving way (cf. Basel Convention [5]).

The means to reach these goals are manifold. In the nineteenth century, “collection and transport of wastes out of town as fast as possible” was the main principle of waste management. Today, sophisticated logistic systems, high-tech mechanical sorting processes, mature and efficient recycling concepts are state of the art in many affluent countries. On the other side, for emerging economies, the economic situation does not allow to apply sophisticated waste treatment processes such as sorting, incineration, or mechanical-biological treatment.

It should be kept in mind that most societies spend between 0.2% and 0.4% of their GDP for waste management [6]. Thus, in a wealthy region with a GDP above US \$ 30.000 per capita, people are willing and can afford to spend about US\$100 per capita annually for waste management. This enables the region to implement modern waste management systems including waste-to-energy. By contrast, in an emerging economy with a per capita GDP below US\$3.000, the financial resources available for waste management are less than 10 US\$ per capita and year. It is clear that for this sum, it is neither possible to construct nor to operate an MSW incinerator. Hence, subsidizing the investment of an MSW incinerator or any other expensive treatment technology in an emerging economy is not an appropriate means to solve waste management problems.

The goals of MSW incineration are congruent with, but more specific than, the goals of waste management and can be summarized as follows:

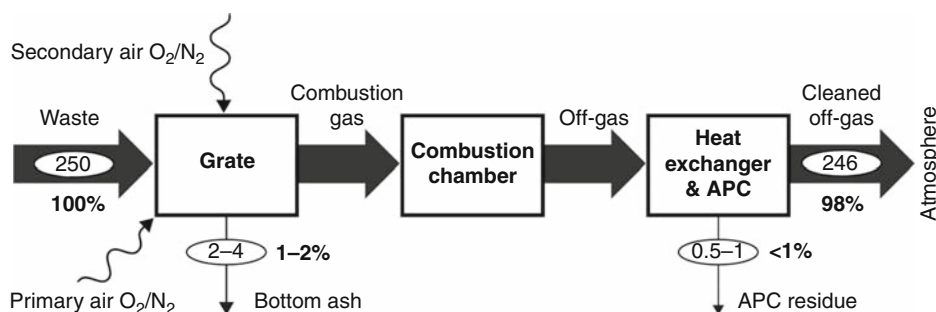
Protection of Humans and the Environment

Disinfection: The original objective of waste incineration is still one of the main – but often forgotten – goals and advantages of modern incineration. No other waste treatment process such as landfilling or composting has a similar hygienic effect. The high temperature and long residence time in modern incinerators guarantees complete destruction of all microorganisms. Thus, incinerators are a necessary final sink for all those materials that are biologically contaminated. This is especially important for high-density megacities, where the outbreak of diseases can have disastrous consequences. A particular case is represented by hospital wastes that require special attention due to their high concentration of pathogenic organisms. Such materials demand special care for

delivery, storage, and feeding. Thus, they need to be packaged and handled with special care when they are co-combusted with MSW, or they are treated in specialized furnaces. In case of an outbreak of contagious diseases, it is of great value to have incinerators at hand that are designed to safely dispose of biologically contaminated materials.

Complete oxidation: A main goal of incineration is to completely mineralize organic substances to CO_2 , H_2O , and other inorganic compounds. As discussed in the previous section, thousands of tons of hazardous organic substances are produced in industry, utilized by consumers, and collected by waste management. The hazardous materials that are built into consumer products and investment goods can neither be recycled nor composted or landfilled because of their persistent and toxic properties. They need to be broken down by thermal treatment at elevated temperatures, with sufficient oxygen supply and at long enough residence times to completely destroy any organic material. If special care is taken, “de novo” synthesis of trace organic substances in the furnace and subsequent boiler and flue gas cleaning can be minimized, and the overall conversion rate for total organic carbon to carbon dioxide and carbonates (bottom ash and APC residue) is close to 99%. The remaining carbon in the WTE bottom ash has been found to consist of 19–38% organic carbon, 33–62% of elemental carbon, and 10–38% of inorganic carbon [7] (Fig. 1).

Emissions compatible with environmental standards: The first goal of waste incineration is to protect the environment from pathogenic organisms, hazardous organic chemicals, and heavy metals and other toxic substances. As will be discussed in section “Goals of Waste Management and of Thermal Waste Treatment”, state-of-the-art incineration has reached this goal. Table 3 presents the example of cadmium on a regional scale: Let us consider a region of 250 km² with a population of 100,000, each producing about 400 kg of MSW per year containing 10 mg Cd/kg MSW. Table 3 shows the increase of Cd in the regional soil, assuming that all emitted Cd settles in the region (such a hypothesis is justified if the region is surrounded by similar regions that import the same amount of Cd as they are exporting). Cadmium is chosen here because in the 1970s, this volatile heavy metal was recognized as the most significant pollutant emitted by MSW



Thermal Treatment of Waste: Key Element for Sustainable Waste Management. Figure 1

Flow of carbon through an MSW incinerator: Carbon is effectively oxidized to CO_2 ; the small amounts of carbon in bottom ash and APC residue are mostly inorganic carbon

Thermal Treatment of Waste: Key Element for Sustainable Waste Management. Table 3 Contribution of modern waste-to-energy plants to cadmium loading of soils

Transfer coefficient, k_{Cd} cleaned off-gas	Emission, mg Cd/Nm^3 of stack gas	Cd flow to soil in 1,000 years, mg Cd/m^2
0.1	0.2	200
0.01	0.02	20
0.001	0.002	2
best available technology		
0.0005	0.001	1
Geogenic background concentration in soils:		90 mg/m^3

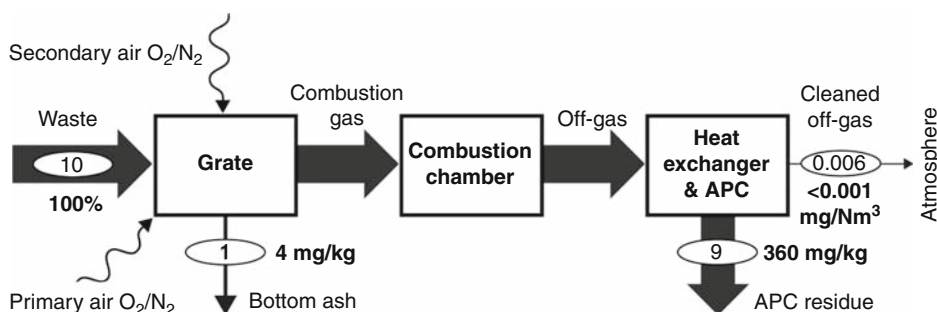
incinerators at that time [9, 10]. The fraction of input Cd that is emitted in the stack gas after the Air Pollution Control (APC) system is expressed by the transfer coefficient. It should be noted that state-of-the-art incinerators operate at transfer coefficients for cadmium of less than 0.0005; as shown in Table 3, emissions from such incinerators are insignificant when compared to other sources, and their contribution to geogenic values of cadmium is low even when long time periods are considered (Table 3).

Concentration of hazardous substances: As shown in Fig. 2, thermal treatment processes have the power to concentrate or deplete elements in the various incineration products. Atmosphiliic metals with comparatively high vapor pressures and low boiling points, such as mercury, cadmium, tin, zinc, antimony and lead, are

preferentially transferred from the combustion grate to the gas phase. During gas cooling, the metal vapors condense on available surfaces which are in general provided by the large number of small particles formed by incineration. These particles, overproportionally loaded with atmosphiliic metals, are removed from the process gas stream by electrostatic precipitators or fabric filters, yielding a filter residue that is highly concentrated in useful but also toxic heavy metals. The overall accumulation effect is remarkable: Starting with 1 Mg (i.e., 1 t) of MSW containing 10 mg of Cd, an MSW incinerator produces about 20–30 kg of APC residue containing 360 mg Cd/kg.

The concentration in the APC residue (also called “fly ash”) during the incineration process occurs only with atmosphiliic metals and chemical compounds with a high vapor pressure, such as metal chlorides. Metals and metal compounds with low volatility such as iron, chromium, titanium, and cobalt accumulate rather in the bottom ash. Due to the different behavior of various metals, it will never be possible to direct all useful, or all hazardous, materials to one particular fraction of incineration products. However, the partitioning described above allows the recovery of some metals from the bottom ash and some from the APC residue.

Immobilization: Bottom ash and APC residues contain soluble compounds such as chlorides, nitrates, or sulfates of metals. Thus, a main concern is the long-term mobility of heavy metals. Due to the individual partitioning of substances, the solubility of bottom ash and APC residue is not the same. Volatile metals and soluble chlorides are mostly transferred to the fly ash product; the bottom ash contains more oxides, and



Thermal Treatment of Waste: Key Element for Sustainable Waste Management. Figure 2

Flow of cadmium through an MSW incinerator; the concentration of cadmium in APC residues increases 40-fold from 10 mg Cd/kg in MSW to 360 mg Cd/kg; the cadmium concentration in bottom ash is 2.5 times lower than in MSW

some sulfates that are less soluble. Both products react with water and air (O_2 , CO_2), and are by no means in a thermodynamic equilibrium when exposed to the environment. The long-term mobility of heavy metals in bottom ash depends on pH and thus on the acid-neutralizing capacity [8]. Bottom ash appears to be predominantly buffered by calcium minerals. Under alkaline conditions, calcium hydroxides/silicates and $CaCO_3$ can dissolve. In neutral-to-acidic conditions, silicate dissolution becomes increasingly relevant. Basic calcium hydroxide/silicate components are likely to be neutralized by CO_2 or silica. The solution chemistry of the APC residues is similarly controlled by pH. Thus, solidification by cement stabilization or other treatment requires careful control of pH conditions, anticipating the future fate of the landfilled material (leaching and environmental conditions).

Resource Conservation

Volume reduction: This is particularly for large cities where landfill space close to waste producers is rare. Incineration decreases the mass of MSW by 70–75% and the volume by 90%. Thus, the lifetime of landfills can be reduced at least by a factor of 4. If materials are further recovered from bottom ash, it is feasible to reduce the required volume for ash disposal to less than 10% of the volume required for landfilling of untreated MSW.

Energy recovery: While initially the recovery of heat was more directed toward the protection of the air pollution control (APC) devices, most of today's MSW incinerators are well embedded in regional

electricity grids and, sometimes, are also linked to district heating networks. In 2009, MSW incinerators across Europe had an annual capacity of 73×10^6 Mg and supplied 30 TWh of electricity and 55 TWh of heat. The share of waste-derived energy in Europe amounts to less than 1% of total energy demand. About 50% of energy derived from MSW combustion originates from renewable sources, such as wood, paper, and biomass. The contribution of MSW incinerators to renewable energy supply in Europe totals about 4%.

Recovery of materials: The products of incineration contain not only hazardous substances but also valuable resources. The recovery of iron from the bottom ash has a long tradition; also, the reuse of bottom ash as construction material has been practiced for hundred years by now (although with limited success). Today, it seems feasible and economically attractive to recover more metals, such as copper, aluminum, and others. Also, the enriched concentration of atmophilic metals in filter ashes potentially opens the possibility to recover metals such as cadmium, zinc, antimony, and others from APC residues. The main objective must be to enrich valuable constituents in highly concentrated small fractions by optimizing the thermal treatment process and, also, to divert those elements that reduce the effectiveness of recovery to a small fraction that can be landfilled. The ideal of zero solid incineration residues to landfill will be difficult to reach, considering that the intricate composition of bottom and filter ash reflects the complex mixture of nearly all elements of the periodic table that are present in MSW.

In view of the goals of waste management and of incineration, what is the contribution of waste

incineration to attaining the goals of waste management? In particular, what is the main contribution of MSW incineration for the protection of humans and the environment, and for resource conservation? These questions will be discussed in the following section.

Thermal Waste Treatment for Environmental Protection

In addition to fulfilling the goals of recycling, waste management must also provide a solution for those materials that cannot be recycled. This applies in particular to the large amount of hazardous substances that are used today in modern products such as biocides, flame retardants, heavy metals, nanoparticles, etc. Mechanical or biochemical treatment such as sorting combined with composting or anaerobic digestion can neither decompose nor eliminate these substances. The key role of thermal treatment is to completely mineralize organic substances, and to control inorganic substances by concentration. At present, no other waste treatment process is available that fulfills this task at the same proven reliability, cost-effectiveness, and environmental soundness.

Toward the end of the nineteenth century, waste incineration was introduced as a means to prevent pollution. The first plants were been built in England and Germany, as an alternative to waste disposal on agricultural land that spoiled valuable farmlands with disease-spreading microorganisms. These incinerators, directed toward the first goal of waste management, did have a stack but they were not equipped with APC systems and did not recover the chemical energy contained in MSW. Their sole purpose was to prevent the application of MSW on valuable farmland by reducing volume and mass. It is interesting to note that the bottom ash of some of these early incineration plants was used as a construction material as far back as the beginning of the twentieth century. Unfortunately, there are no records as to why this practice was abandoned later; it is likely that the bricks made at that early time from incinerator ash did not fulfill their purpose in the long run.

The next step of development was the introduction of simple filters for the removal of particulates, and of energy recovery. Both went hand in hand, since even simple filter systems require the cooling of flue gas temperatures. Municipal incinerator plants built in the late

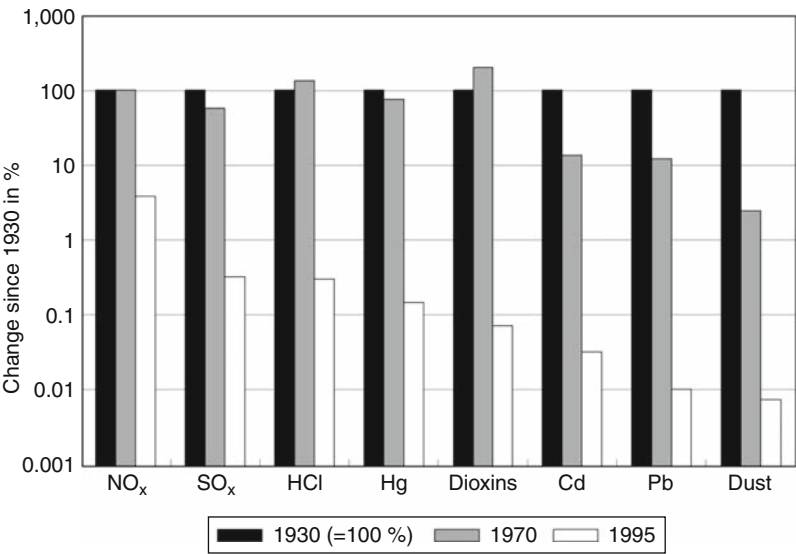
1960s were already equipped with sophisticated energy recovery (combined district heating with electricity production), but the air pollution control devices removed only coarse particulate matter (legal standards at that time were around 150 mg of dust per Nm³ off-gas). Waste incineration got a boost during the oil crisis of 1973 when fuel prices skyrocketed. Shortly after, the first groundbreaking studies on heavy metals in incinerator off-gas appeared in the scientific literature [9, 10]. This had little effect until the 1980s when – under heavy pressure from an informed public and legislation in countries like Switzerland, Denmark, and Sweden – the waste-to-energy industry developed highly efficient APC pollution control devices.

Today, emissions from state-of-the-art waste incinerators are among the lowest of all industrial power plants. Thanks to technological advancements that were inconceivable in the 1970s, modern municipal solid waste incinerators can beat most stringent existing regulations easily, for some parameters even by orders of magnitude. Figure 3 summarizes the reduction emissions from municipal solid waste incineration from 1930 to 1970 and 1995. The graph stops at 1995 because the highly advanced state of the art of APC of 1995 has not changed much since then; further improvements have focused on energy efficiency, cost-effectiveness, and reliability.

Particulate matter and lead have been reduced by about four orders of magnitude, and most other heavy metals and dioxins by three orders of magnitude. Also, hydrochloric acid and sulfur oxides have been drastically eliminated, and only for nitrogen oxides the reduction was somewhat moderate. The increase in HCl and dioxins in the 1970s was due to the introduction of PVC that made up about 1% of the mass of MSW.

Table 4 shows that today – in contrast to some earlier publications – the share of MSW incineration for total dioxin emissions is insignificant. Of course this applies only for countries employing BAT (in the USA: MACT) incineration, such as Austria, Germany, Japan, the Netherlands, Singapore, Sweden, Switzerland, and the USA.

In view of the objectives of waste management, it is important to point out the pollution potential by burning household waste in private stoves. This practice might appear attractive in areas with high penetration of solid fuel household furnaces. Such furnaces, designed to



Thermal Treatment of Waste: Key Element for Sustainable Waste Management. Figure 3
Development of emissions from municipal solid waste incineration from 1930 to 1995 on a concentration base (mass per volume of off-gas)

Thermal Treatment of Waste: Key Element for Sustainable Waste Management. Table 4 Contribution of MSW incineration to national dioxin/furan load in Germany and Austria (calculation is based on data from [11] and [12], assuming that all MSW generated is incinerated)

Country	Total emissions of PCDD/PCDF (g TEQ/year)	Share of MSW incineration	
		g TEQ/year	%
Germany	800–1,200	5.5	0.5–0.7
Austria	50–320	0.53	0.2–1.1

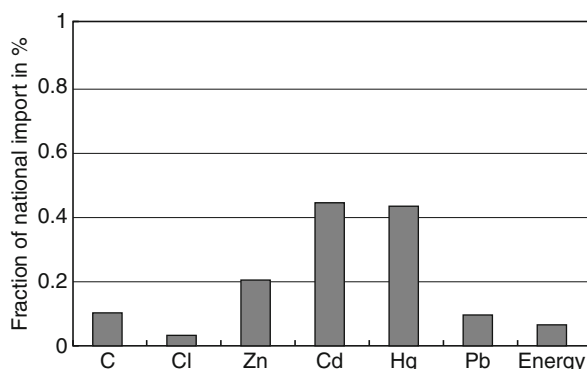
produce heat without generating greenhouse gases, are becoming increasingly popular, and are sometimes even subsidized by authorities. With increasing fuel costs, some waste will eventually be used as a fuel for such furnaces, resulting in severe air pollution.

Figure 3 can be used to compare emission values of today’s MSW incinerators (emissions of 1995) and household furnaces (emissions of 1970). It is striking that even if only 1% of all waste is incinerated in household furnaces, the emission load of dioxins and particulates will be much larger than for state-of-the-art MSW incineration. It is important to convey this

message to the general public that aims at reducing greenhouse gases by burning waste-derived products in their heating systems.

MSW is more important as a carrier of heavy metals than as an energy carrier (Fig. 4). Thus, it is important to stress that incineration is first of all a process that protects the environment from harmful organic and inorganic substances. The generation of useful energy is an added value, but not the main purpose. This is important when it comes to the disposal or use of bottom ash and filter ash. If a posttreatment of incineration residues such as concentration, solidification, or immobilization requires energy, this should not be weighed against the benefit of protecting the environment from harmful substances. After all, the turnover of energy in waste management is marginal when compared to total energy demand. Today, the challenge of producing incineration residues that have “final storage quality” has been recognized and is in focus of research and development. While treated bottom ash is close to final storage quality, in Germany and some other countries, the APC residues are disposed of in underground salt mines where the risk of water contact and thus mobilization is exceptionally small.

In terms of greenhouse gas reduction, incineration has a double advantage: Since about 50% of MSW is of



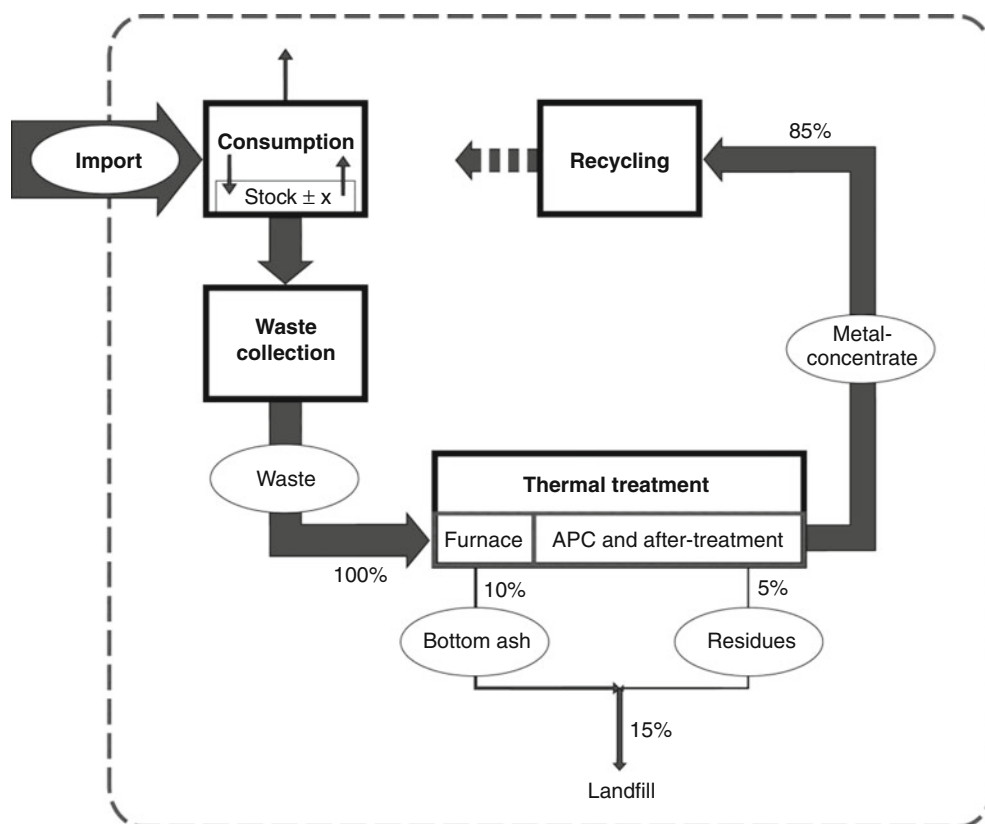
Thermal Treatment of Waste: Key Element for Sustainable Waste Management. Figure 4

Fraction of selected substances and energy contained in MSW compared to national import flow: the case of Switzerland 1990–2000

biogenic origin [13], it contributes also to the reduction of greenhouse gases by replacing fossil fuel derived energy. Also, in contrast to composting producing “cold” CO₂ or landfilling with limited gas recovery, the organic carbon is not released as unproductive CO₂ and CH₄ to the atmosphere but is used to produce electricity and/or heat.

The Contribution of Thermal Waste Treatment to Resource Conservation

As a result of advanced APC technologies, MSW incinerators have become marginal polluters of air, water, and soil when compared to other high-temperature industrial processes. At the same time, as waste-to-energy capacity increases, the amount of solid incineration products, in particular APC residues, is also



Thermal Treatment of Waste: Key Element for Sustainable Waste Management. Figure 5

Recovery of atmophilic metals from APC residues allows substantial recycling of particular metals such as cadmium, zinc, tin, and antimony. Taking into account that MSW contains nearly half of the national import of Cd, about 40% of this metal used could be recycled by recovery from the APC residues of MSW incineration

increasing and poses a new challenge. In the future, main emphasis will be laid on beneficial use and/or disposal of the solid products of incineration, namely, bottom ash and APC residues. Both materials comprise considerable amounts of valuable substances in concentrated form. While iron scrap in bottom ash is already routinely collected and economically recycled, current research focuses more on the reclamation of aluminum, copper, and other nonferrous metals. It appears that substantial amounts can be recovered at competitive costs. The goal will be to separate those materials that are clean enough and well suited for construction purposes (stones, sand, concrete, and other residues from construction materials).

Certain atmophilic heavy metals such as cadmium are highly enriched in APC residues. Therefore, this material presents not only a risk but also an opportunity for resource recovery. About 50% of the national import of Cd ends up finally in MSW and similar wastes that are suitable for MSW incineration (Fig. 4). Roughly 80% of the Cd entering an incinerator leaves the plant as APC residue (the partitioning between bottom ash and APC residues depends upon incineration design, operating conditions, and waste composition, e.g., Cd in batteries or in plastic additives). Therefore, as shown in Fig. 5, nearly half of the Cd used in a region can be recovered from the APC residue of waste-to-energy plants within the region. Similar material recovery is possible for other atmophilic metals such as Zn, Pb, Sn, and Sb. However, for most of these other metals, the ratio of metal in MSW to national imports is less than 50% (Fig. 4).

Future Directions

The potential of MSW incineration for improving waste management is not yet fully exploited. The following outlook summarizes some of the ideas that have been presented recently, again listed in the order of: first, health and environmental protection and, second, resource recovery.

While average emissions are exceptionally low, some emission peaks can be discerned when short-term time series are analyzed. Mitigation measures such as waste pretreatment, in situ homogenization, or alternative thermal treatment technologies should be examined. Since new materials such as nanoparticles and rare metals

are penetrating the market, the fate of these new substances during combustion as well as in residues resulting from incineration needs to be investigated.

MSW incinerators can also be used as tools to monitor MSW composition. On one hand, biogenic carbon can be distinguished from fossil carbon. As stated earlier, incineration contributes toward the reduction of greenhouse gas emissions. Several methods exist to measure the biogenic fraction of CO₂ in the off-gas of incineration with good accuracy [13]. In order to take advantage of the greenhouse gas reduction potential of MSW incineration, incinerators should be included in emission trading schemes.

Incinerators can also be used for cost-effective routine analysis of elemental waste composition. For this purpose, an initial mass balance for the determination of transfer coefficients is made. It is then followed by identifying those incineration products that are the main carrier of certain substances, such as stack gas for carbon, scrubber water for chlorine, APC residue dust for cadmium and antimony, and bottom ash for aluminum and iron. On the basis of the transfer coefficients, the selected elements can be easily determined by measuring the concentration in a well-accessible incineration product of known mass flow, such as the APC residue [14]. The main advantage of this method for waste analysis is that the incinerator acts as a large homogenizer. Consequently, sample size and frequency can be much smaller as compared to direct characterization of the MSW and the results will be more reliable than those from classical waste analysis.

Concerning resource recovery, large-scale research programs have been started in order to recover additional resources from bottom ash [15]. So far, these studies of sophisticated mechanical separation processes point out that many more valuable substances can be extracted from bottom ash. Some of the proponents go as far as to state that, in the future, separate collection of some waste fractions will become obsolete due to the new possibilities of recovering selected substances from incineration products [16].

Concerning APC residues, new concepts are still in the visionary phase. Since these materials make up only a small fraction of the initial MSW (2–3%), it seems feasible to collect them for central treatment at a few regional centers, e.g., serving a population of 10 million.

The resulting amount of 100.000 Mg of APC residue would represent a resource base that would allow for economic recovery of certain metals.

In the future, new methods will be required for evaluation of waste management scenarios. The present LCA approach is not well suited to assessing long-term challenges such as landfill emissions during the next five centuries, or multiple recycling of hazardous substances in recovery schemes. New evaluation criteria such as the statistical entropy analysis SEA [17] (see contribution by Rechberger in this volume) or the final sink approach [18] represent additional means for an objective appraisal of MSW incineration.

MSW incineration has gone a long way from simple combustion to sophisticated waste-to-energy (WTE). It will certainly keep on advancing in the future. In order to conquer and penetrate new markets, it will be necessary to decrease costs of investment, operation, and maintenance. The art will be to deliver the same performance regarding environmental protection, resource conservation, and reliability at lower costs. Considering that waste incineration is now a mature business, with decades of solid and successful experience, this new goal seems feasible. It will ensure that a larger percentage of the global population can enjoy the benefits of this key waste treatment technology.

Bibliography

- Baccini P, Brunner PH (in press) *Metabolism of the anthroposphere – analysis, evaluation and design*. MIT Press, Cambridge, MA
- Brunner PH (2004) Material flow analysis and the ultimate sink. *J Ind Ecol* 8(3):4–7
- Tarr AJ (1997) *The search for the ultimate sink – urban pollution in historical perspective*. The University of Akron Press, Akron
- Belevi H, Baccini P (1989) Long-term behavior of municipal solid waste landfills. *Waste Manage Res* 7(1):43–56
- Basel Convention (1992) Retrieved 27 Jan 2011 from <http://www.basel.int/>
- Brunner PH, Fellner J (2007) Setting priorities for waste management strategies in developing countries. *Waste Manage Res* 25:234–240
- van Zomerena A, Comans RNJ (2009) Carbon speciation in municipal solid waste incinerator (MSWI) bottom ash in relation to facilitated metal leaching. *Waste Manage* 29(7):2059–2064
- Johnson CA, Brandenberger S, Baccini P (1995) Acid neutralizing capacity of municipal waste incinerator bottom ash. *Environ Sci Technol* 29(1):142–147
- Greenberg RR, Zoller WH, Gordon GE (1978) Composition and size distribution of particles released in refuse incineration. *Environ Sci Technol* 12:566–573
- Greenberg RR, Gordon GE, Zoller WH, Jacho RB, Neundorff DW, Yost KJ (1978) Composition of particles emitted from the Nicosia municipal incinerator. *Environ Sci Technol* 12:1329–1332
- Orthofer R, Vesely A (1990) Abschätzung von toxischen Emissionen (PCDD, PCDF, PAH, BaP) aus Verbrennungsprozessen in Österreich (Estimation of toxic emissions (PCDD, PCDF, PAH, BaP) from combustion processes in Austria), Seibersdorf Research Report OEFZS 4554
- Wintermeyer D, Rotard W (1994) Dioxin Emissionen in der BRD – Versuch einer Bilanzierung. *Staub Reinhaltung Luft* 54: 81–86
- Fellner J, Cencic O, Rechberger H (2007) A new method to determine the ratio of electricity production from fossil and biogenic sources in waste-to-energy plants. *Environ Sci Technol* 41(7):2579–2586
- Morf L, Brunner PH (1998) The MSW Incinerator as a monitoring tool for waste management. *Environ Sci Technol* 32(12): 1825–1831
- ZAR (2011) Stiftung zentrum für nachhaltige ressourcen- und abfallnutzung. Retrieved 31 Jan 2011 from <http://www.zar-ch.ch/>
- Bunge R (2008) *Metalle aus Abfall: geld stinkt nicht*, Innovationsworkshop IPEK 02/2008 University of Applied Sciences Rapperswil, Switzerland
- Rechberger H, Brunner PH (2002) A new, entropy based method to support waste and resource management decisions. *Environ Sci Technol* 34(4):809–816
- Döberl G, Brunner PH (2004) Substances and their final sinks – a new indicator for monitoring sustainability. Symposium indicators for evaluating sustainable development – the ecological dimension, Berlin, Germany, 1–2 Nov 2004

3D Pose Estimation of Vehicles Using Stereo Camera

BJÖRN BARROIS¹, CHRISTIAN WÖHLER²

¹EvoBus GmbH, Neu-Ulm, Germany

²Image Analysis Group, Dortmund University of Technology, Dortmund, Germany

Article Outline

Glossary

Definition of the Subject

Introduction and State of the Art

Discussion

A System for Stereo-Based 3D Pose Estimation of Vehicles

Experimental Evaluation
 Summary and Conclusion
 Future Directions
 Bibliography

Glossary

3D Pose estimation Estimation of the translational and rotational degrees of freedom of an object, usually corresponding to its position in 3D space and its orientation angles. For articulated objects additional, internal degrees of freedom are estimated.

Driver assistance system A system which helps the driver of a car in specific traffic situations in order to increase safety.

Iterative closest point (ICP) algorithm An algorithm which finds a set of translation and rotation parameters fitting a measured point cloud to a model.

Optical flow The optical flow vector denotes the motion in image space of the projection of a scene point between two subsequent images of a sequence.

Quaternion Number system which extends the system of complex numbers. In computer vision and computer graphics, quaternions are used for the efficient representation of rotations.

Scene flow Vector field describing the 3D scene points visible in the image and their current motion in 3D space. Projecting the scene flow into the image plane yields the optical flow.

Stereo vision 3D reconstruction of a scene based on a pair of images.

Yaw angle Angle measuring the rotation of a vehicle around its vertical axis.

Yaw rate Temporal derivative of the yaw angle.

Definition of the Subject

For advanced driver assistance systems, the 3D poses and motion states of oncoming and intersecting vehicles represent important information. This work describes methods for 3D vehicle pose estimation based on a motion-attributed 3D point cloud generated. First, stereo and optical flow information is computed for the investigated scene. A four-dimensional clustering approach separates the static from the moving objects in the scene. The iterative closest point

algorithm (ICP) estimates the vehicle pose using a cuboid as a weak vehicle model. Classical ICP optimization is based on the Euclidean distance metric. Its computational efficiency can be significantly increased by applying a quaternion-based optimization scheme. In vehicle-based small-baseline stereo systems, it is favorable to use a polar distance metric which especially takes into account the error distribution of the stereo measurement process. To derive the pose parameters and simultaneously their temporal derivatives, i.e., the motion state of the vehicle, from two subsequent stereo image pairs in order to avoid a temporal filtering stage, a model-based scene flow method is applied. Here, 3D points on the object surface are reprojected into the stereo images and similarities between image windows extracted around the reprojected points are maximized. An experimental evaluation is performed on seven different real-world sequences, where different stereo algorithms, baseline distances, distance metrics, and optimization algorithms are examined. The results show that the proposed polar distance metric yields a higher accuracy especially for the estimation of the yaw angle of a vehicle than the common Euclidean distance metric, especially when using pixel-accurate stereo points. The model-based scene flow approach yields a refined 3D pose estimation along with the instantaneous motion state of the vehicle.

Introduction and State of the Art

Some approaches for detection of obstacles around the ego-vehicle can be found in the literature, but the existence of an obstacle is not enough information to evaluate the situation entirely. The knowledge about pose and motion of oncoming and intersecting vehicles is important to avoid accidents. Especially at intersections, analysis of the behavior of other road users is important to help the driver to pass through the intersection in a safe and comfortable manner. Furthermore, pose information is important for the recognition and prediction of driving maneuvers and the analysis of complex traffic situations involving several traffic participants. This section provides an overview of the state of the art in the field of vehicle pose estimation, where approaches relying on different sensor concepts are discussed.

Systems Based on Monocular Images

A system for segmentation and detection of vehicles using a monocular camera system is described in [1], where a vehicle model defined by seven parameters is used. The position of the vehicle is determined by taking into account the symmetry of the edges in the image, the shape of the model, and the shadow of the vehicle. An evaluation of this approach is not available.

In [2] a Kalman filter-based system for tracking vehicles in front of the ego-vehicle is presented. The contours of other vehicle candidates are extracted using a Hough transform and circumscribed by a bounding box. Furthermore, a classifier determines if the bounding box corresponds to a vehicle or a background pattern. A 3D model is used for tracking the vehicles in 3D space, where the measurement is interpreted as the projection of the 3D model into the 2D image plane, which yields the 3D position of the vehicle. System evaluation is performed based on radar data, where an accuracy of the distance estimation of the image-based approach of better than 1 m at a distance of 60 m is achieved.

Early methods for vehicle detection in the field of traffic surveillance are proposed in [3–5]. These approaches are based on observing a road intersection with a fixed camera. The vehicles are represented by polyhedral models which are adapted to the edges extracted from the images. A 3D scene model provides information about free space and possible occlusions. An illumination model is used to suppress the influence of shadows on the pose estimation. To obtain a prediction of the motion behavior, a description of the possible motion patterns based on a physical model is used. The vehicle is tracked across the sequence using optical flow analysis combined with an extended Kalman filter.

Systems Based on Sensor Fusion

A system consisting of a camera and a radar sensor is used in [6] to detect traffic participants. The advantages of the two sensors (high lateral resolution of the camera, high accuracy of the distances and radial velocities measured by the radar sensor) are used to estimate the position of vehicles driving ahead of the ego-vehicle. The distance is measured based on radar information while the lateral extension of the object is

inferred from the camera image. The evaluation of the system is restricted to the accuracy of the estimated bounding box in the image.

The combination of a monocular color camera and a radar sensor is proposed in [7]. Shadows, rear lights, line segments, and symmetric features are extracted from the image and combined with the radar information in a low-level fusion step. A subsequent post-processing step aims for removing false detections and determining the vehicle positions. An evaluation of the metric accuracy of the system is not available.

A laser scanner and a video camera are used in [8] to estimate the pose, size, and velocity of vehicles. The detection and 3D pose estimation of vehicles is obtained based on the laser scanner data. Relying on the pose estimation result, rectified side views of the detected vehicles are constructed, which are in turn used by a cascade classifier to distinguish vehicles in front of the ego-vehicle from background patterns. The evaluation in [8] is limited to the classification performance of the system, while the geometric detection accuracy is not examined.

Similarly, a combination of a laser scanner and a camera is used in [9] for the recognition of objects in traffic scenes. This approach is based on sensor fusion relying on vehicle-specific features extracted from both sensors. Besides the application of autonomously following the vehicle driving ahead, traffic jam and road intersection scenarios are examined. It is shown that the proposed system is able to track vehicles even in the presence of high accelerations.

Stereo Camera Systems

The early approach to stereo-based vehicle detection described in [10] relies on a stereo camera system with a relatively long baseline approximately corresponding to the width of the ego-vehicle. Vehicles driving behind the ego-vehicle are extracted from the 3D point cloud, and a Kalman filter is applied, where the highway scenario is addressed specifically. Due to the long baseline, an estimation of the vehicle position is possible for distances of 100 m and more.

Stereo information and optical flow data are combined in [11] to determine the position and the velocity of vehicles. The 3D points and the extracted optical flow field are processed simultaneously by an extended

Kalman filter framework. Objects are inferred from the scene along with their velocities using a clustering technique. A considerable settling time of the temporal filtering stage is observed. The accuracy of the proposed method is not evaluated.

A fairly complex system relying on 3D scene reconstruction based on the fusion of structure from motion data and stereo vision is described in [12]. A detection step yields the objects present in the scene along with their approximate orientation in discrete steps and an assignment to different object classes using the approach introduced in [13]. The results are reliable in a depth range between 15 and 30 m. The pose of the objects is not compared to ground truth data.

A stereo camera is used in [14] to detect vehicles driving ahead and estimate their distance to the ego-vehicle. In a first step, a stereo-based 3D reconstruction of the scene is performed. A symmetry analysis extracts rear views of vehicles from one of the camera images, and a bounding box is computed. The vehicle is then tracked in the 2D image space and associated with the 3D data in a separate step.

Stereo analysis based on affine warping is used in [15] to detect obstacles ahead of the ego-vehicle. The image of the left camera is warped and mapped onto the image of the right camera using affine warping and edge comparison. Dynamic programming estimates a boundary between the road surface and obstacles on the road. However, the orientations of the obstacles are not estimated.

In [16], a stereo camera system is used for several applications, e.g., the detection and tracking of vehicles and pedestrians. An occupancy grid distinguishing between free and occupied regions in front of the vehicle is created using the 3D points of an initial stereo analysis to detect objects in front of the car. Once an object is found, its size, aspect ratio, and orientation are used to find a corresponding model in an object model database. The most suitable 3D model from the database is projected into the image and fitted to the image of the object using chamfer matching. The distance up to which this approach yields reliable results is about 35 m. No evaluation of the pose estimation accuracy is given.

The approach introduced in [17] uses a sparse scene flow field which is established by a point-wise Kalman filter-based integration of stereo and optical flow

information. An extended Kalman filter tracks the point cluster representing the vehicle. The resulting motion-attributed 3D point cloud is segmented according to their position and motion state, where the independently moving parts of the scene are regarded as objects. The objects are tracked using a second, object-specific Kalman filter stage. An explicit vehicle model is not utilized, but the size of the vehicle is estimated based on the size of the point cloud. The proposed system is evaluated on synthetically rendered image sequences. The settling time of the Kalman filter framework typically amounts to about 25 frames (cf. also [11]). An extension of this approach is proposed in [18], where the observed object trajectories are analyzed by a particle filter stage in order to obtain a prediction of the vehicle motion for time intervals of several seconds. Example images and pose estimation results are shown in Fig. 1.

Three approaches to the adaptation of a temporal filtering stage are discussed in [19] for vehicle tracking based on stereo information. Specifically, an interacting multiple model (IMM), a maximum likelihood estimation of the filter parameters based on the yaw rate, and an extended model taking into account the yaw acceleration are discussed. The proposed tracking approaches are evaluated using synthetically generated images as well as 3D point clouds synthetically generated based on real trajectories of the vehicle center. On these data sets, the positional accuracy of the system amounts to 0.1–0.2 m and the accuracy of the filtered yaw rate to about 2–6° per second.

Discussion

In many state-of-the-art systems, the orientation angle of the vehicle remains undetermined or needs to be inferred from the specific application scenario, e.g., when vehicles driving ahead of the ego-vehicle are detected. Other systems estimate the orientation angle by a temporal filtering stage, which has the disadvantage of delays occurring due to the finite settling time. For an accurate prediction of the motion behavior the direct measurement of the vehicle orientation is essential. Furthermore, metric evaluations of the estimated poses in comparison to independently determined real-world reference data can hardly be found in the existing literature, such that it is difficult or impossible to



3D Pose Estimation of Vehicles Using Stereo Camera. Figure 1

Tracking results obtained with the stereo-based method described in [17, 18] for an oncoming vehicle first detected at 48 m distance. The predicted driving path indicating curve radius and velocity for the next second and the tracked 3D points are shown as well as a bounding box indicating the object pose (from [18])

compare different existing approaches with respect to measurement accuracy.

A general disadvantage of systems using monocular cameras is the low accuracy of the resulting distance measurements, while systems relying on sensor fusion require a considerable instrumental and calibration effort. In principle, a laser scanner is a sensor which is well suited for measuring the 3D pose of vehicles due to its high lateral resolution and distance accuracy, but laser scanners still tend to be too expensive for commercial automotive systems. Hence, stereo cameras are a promising cost-efficient alternative as they provide both a high lateral resolution and a reasonable accuracy of the measured distances in ranges relevant for typical traffic scenarios.

A System for Stereo-Based 3D Pose Estimation of Vehicles

This section describes approaches for 3D pose estimation of vehicles at high accuracy without relying on temporal filtering. A tracking-by-detection approach is performed where the pose estimation is performed individually for each frame. First, the input images from a calibrated camera system are rectified to standard geometry with epipolar lines parallel to the image rows. A sparse scene flow field is computed to obtain reliable depth information combined with motion information about the investigated scene, where three different algorithm combinations are used to compute

a sparse incomplete scene flow field, i.e., a motion-attributed 3D point cloud. Separate objects are extracted by clustering, where each point is defined by three spatial coordinates and one or two velocity coordinates. The 3D points associated with the object are used by an iterative closest point (ICP) method to determine the object pose, where the classical Euclidean distance or a polar distance metric can be used. In a last step, a model-based scene flow stage yields a refined estimation of the pose parameters and their instantaneous temporal derivatives.

Extraction of Depth and Motion

Spacetime Stereo (adopted from [20] and [36]) The spacetime stereo algorithm is based on local intensity modeling and yields 3D points with the associated motion component parallel to the epipolar lines [20]. Image regions corresponding to a sufficiently high vertical intensity gradient are extracted in the left and right camera image, and their local spatiotemporal neighborhood is modeled by the model function $h(\vec{P}, u, v, t)$ where u and v denote the pixel coordinates, t the time coordinate in a spatiotemporal region of interest, and $\vec{P}$ the vector of function parameters:

$$h(\vec{P}, u, v, t) = p_1(v, t) \tanh [p_2(v, t)u + p_3(v, t)] + p_4(v, t) \quad (1)$$

The tanh function approximates the shape of an ideal edge blurred by the optical system. The functions $p_i(v, t)$ are polynomials in v and t of (in principle) arbitrary order. In the experiments described in section “[Experimental Evaluation](#),” linear polynomials were used. Correspondence analysis is then based on a comparison of the modeled edges in the left and the right image in terms of the sum of squared differences (SSD). The SSD values on each epipolar line are analyzed, where different constraints can be taken into account: uniqueness, ordering, or the minimum weighted matching constraint [21]. Analysis of the model function parameters yields the velocity component along the epipolar line and sub-pixel-accurate disparity values. An example of the intensity profile of an observed moving edge versus a moving edge modeled according to [Eq. 1](#) is shown in [Fig. 2](#).

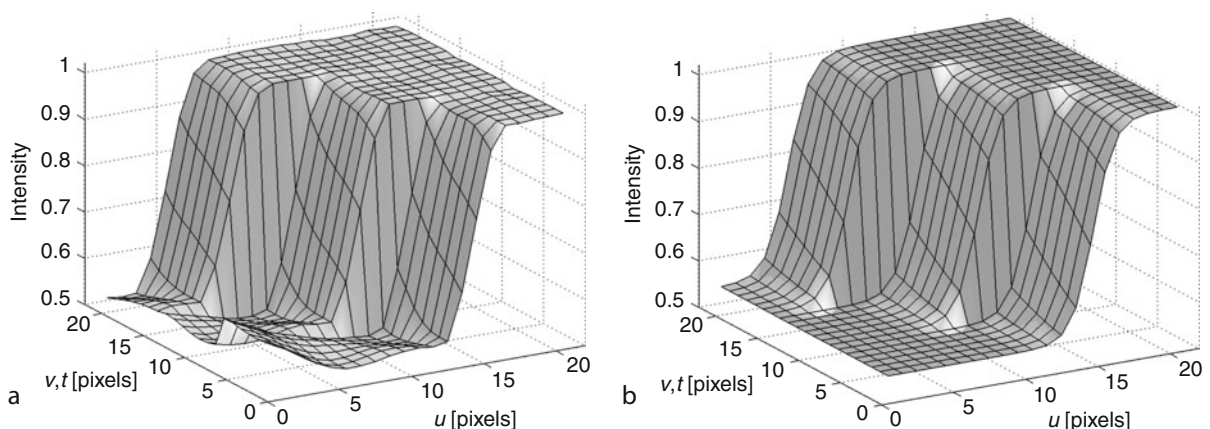
Feature-Based Stereo and Optical Flow (adopted from [36]) The utilized feature-based method for computing stereo and optical flow is based on local features to obtain a representation of the investigated scene [22]. The features are established in both images (left and right camera or current and previous time step), then the comparison is performed using a hash table technique. This approach provides a fast implementation of the algorithm. The correspondences have

pixel accuracy, which is a disadvantage in comparison with the other regarded stereo algorithms. The combination of the stereo and optical flow results yields 3D points associated with the motion parallel to the image plane.

Correlation Stereo and Feature-Based Optical Flow (adopted from [36]) The combination of correlation-based stereo with the feature-based optical flow technique yields optical flow and more precise depth information in video real time. The correlation-based stereo method performs a SSD comparison of left and right image patches [23]. An interest operator (Prewitt filter) is used to separate reliable from unreliable depth information. The combination yields 3D points with two-dimensional motion vectors.

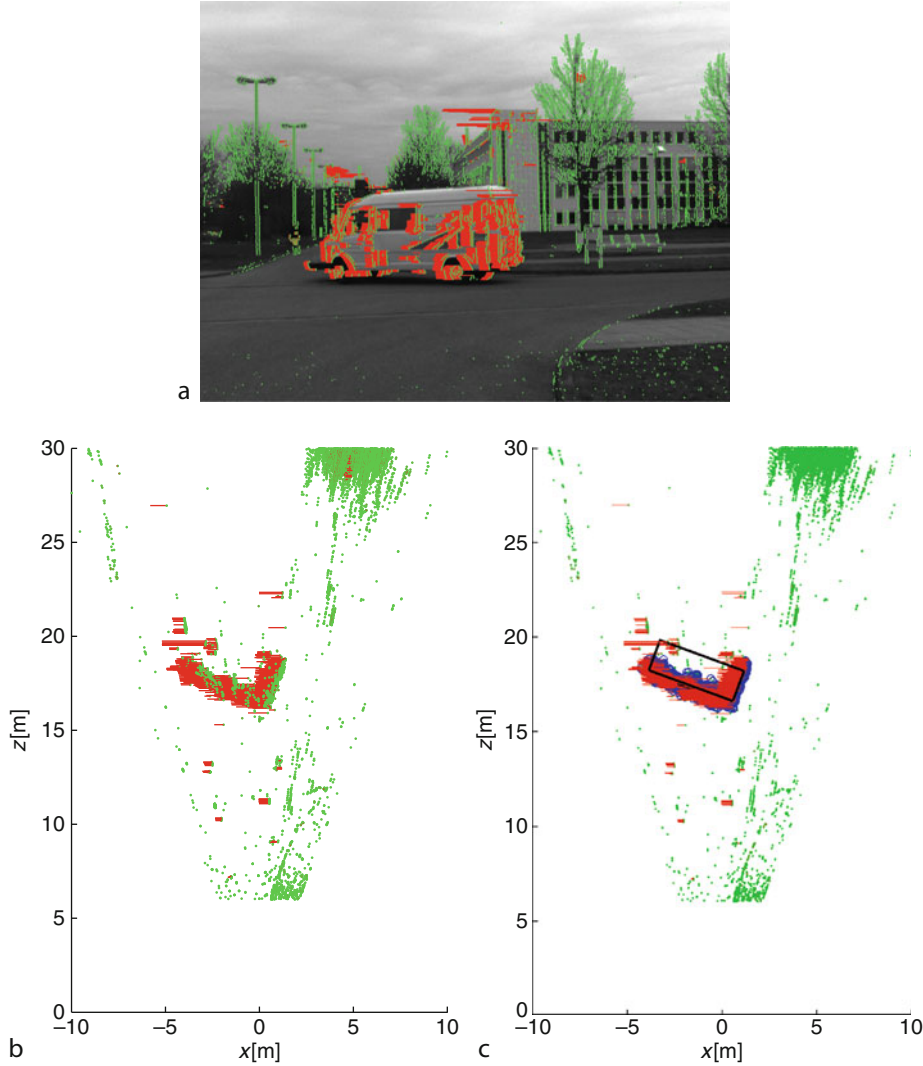
Clustering

This section provides an extended description of the clustering technique outlined in [36]. An initial segmentation of the attributed 3D point cloud extracted with any of the sparse scene flow techniques (cf. [Fig. 3a](#)) is obtained by means of a graph-based unsupervised clustering technique [24] in a space of 4, 5, or 6 dimensions spanned by the three spatial coordinates and the extracted velocity components of the



3D Pose Estimation of Vehicles Using Stereo Camera. Figure 2

(a) Spatiotemporal intensity profile of a moving edge, measured over a time interval of three time steps. The size of the spatiotemporal matching window is $21 \times 7 \times 3$ pixels. For visualization, the (v, t) axis is divided such that each time step comprises an interval of seven pixels. (b) Modeling result according to [Eq. 1](#) (from [37])



3D Pose Estimation of Vehicles Using Stereo Camera. Figure 3

Example of an incomplete sparse scene flow field. (a) Original image with projected 3D points (*green*) and associated motion vectors (*red*). (b) Bird's-eye view of the same scene. (c) Clustered points belonging to the vehicle (*blue*) together with the adapted model (*black*) (adapted from [36])

3D points [25]. This clustering stage generates a scene-dependent number of clusters, essentially separating the moving object from the (stationary or differently moving) background.

To avoid the need for a parameter that weights spatial against velocity coordinates, the graph-based method in [24] is used which takes into account the spatial and velocity coordinates separately. Accordingly, for two scene flow points

$$\vec{P}_1 = \begin{pmatrix} x_1 \\ y_1 \\ z_1 \\ \dot{x}_1 \\ \dot{y}_1 \\ \dot{z}_1 \end{pmatrix} \quad \text{and} \quad \vec{P}_2 = \begin{pmatrix} x_2 \\ y_2 \\ z_2 \\ \dot{x}_2 \\ \dot{y}_2 \\ \dot{z}_2 \end{pmatrix}, \quad (2)$$

the distance measure $\vec{\delta} = (\delta_e, \delta_v)^T$ of the clustering algorithm is a vector with two components given by

$$\begin{aligned}\delta_e &= \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2} \\ \delta_v &= \sqrt{(\dot{x}_1 - \dot{x}_2)^2 + (\dot{y}_1 - \dot{y}_2)^2 + (\dot{z}_1 - \dot{z}_2)^2}.\end{aligned}\quad (3)$$

The clustering approach is based on two thresholds θ_e (for the spatial distance) and θ_v (for differences in velocity). Two scene flow points can be assigned to the same cluster if $\delta_e < \theta_e$ and $\delta_v < \theta_v$. Hence, it is not necessary to combine the different features into a single scalar value as it would be necessary for classical clustering approaches such as the k -means algorithm [26].

The method in [24] relies on a graph-based technique to establish clusters, where the scene flow points are regarded as the nodes and the distances between them as the edges of the graph. The points are sorted according to the norms of their associated velocities, and the resulting sorted list $S = \{\vec{P}_1, \dots, \vec{P}_n\}$ is traversed completely. For each point $\vec{P}_i = (x_i, y_i, z_i, \dot{x}_i, \dot{y}_i, \dot{z}_i)^T$ with $i = 1, \dots, n$ its distances according to Eq. 3 to the subsequent points in the sorted list are computed. If the velocity distance δ_v to one of the subsequent points falls below the threshold θ_v , the search is terminated. If $\delta_e < \theta_e$ and $\delta_v < \theta_v$, the two points are assigned to the same cluster. Once the complete sorted list has been traversed, indices are assigned to the clusters based on a recursive algorithm.

Important advantages of this clustering algorithm in the context of vehicle detection are its high computational efficiency, its simple parameterization, and the fact that the overall number of clusters does not need to be known in advance. The nontrivial problem of appropriately weighting spatial against velocity coordinates does not occur. Since only velocity differences between neighboring points are regarded, the algorithm avoids a decomposition of rotating objects such as turning vehicles into different clusters – as a result of the velocity components varying continuously along the object axis, such a decomposition would occur for many classical clustering approaches which rely on single scalar distance metrics.

Pose Estimation Based on the ICP Algorithm

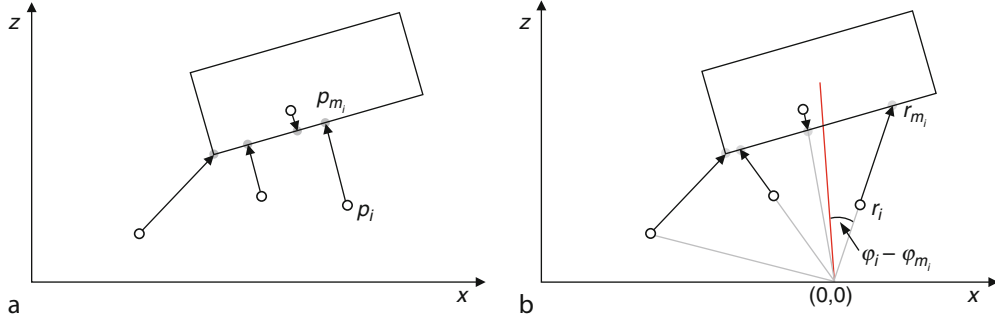
The outline in this section of the ICP algorithm is adopted from [25] and [36]; it is extended by the

description of a computationally efficient quaternion-based optimization approach. A classical method to register a point cloud to a geometric object model is the iterative closest point (ICP) algorithm introduced in [27]. Given an initial estimate of the object pose, the pose parameters are updated by minimizing the mean squared distance between the scene points and the model, and the point assignment is repeated accordingly. This procedure is applied in an iterative manner. As a result, the algorithm yields the three-dimensional object pose. In [27], this method is applied to the registration of point sets, curves, and surfaces. However, the approach can only be used in situations where all scene points belong to the object as it is a pose estimation rather than a scene segmentation technique.

In the ICP algorithm proposed in [28], the scene points as well as the object model are represented as sets of chained points. During each iteration step, the pose parameters are updated, and at the same time some scene points are assigned to the object model while others are rejected, based on the distance to the object and the similarity of the tangent directions of the scene and model curves. Thus, outliers in the point cloud data are automatically rejected, and the algorithm is robust with respect to disappearing and reappearing object parts as well as partial occlusions. As a result, the algorithm yields the subset of scene points belonging to the object, i.e., a scene segmentation, along with the object pose.

Euclidean Distance Metric (adopted from [36]) For pose estimation of vehicles, a cuboid is utilized as a weak geometric model representing several vehicle types. An initial pose is estimated based on the centroid and the first principal component of the vehicle point cloud. Afterward, the ICP algorithm fits the geometric model to the point cloud. Thus, the translational pose parameters t_x and t_z and the yaw angle θ are updated by minimizing the mean squared distance between the scene points and the model. The distance d_i is the perpendicular between the visible model side and the 3D point in Euclidean space (cf. Fig. 4a), where $\vec{p}_i$ denotes the 3D point and $\vec{p}_{m_i}$ the corresponding point on the model plane:

$$d_i(t_x, t_z, \theta) = \|\vec{p}_i - \vec{p}_{m_i}\|. \quad (4)$$



3D Pose Estimation of Vehicles Using Stereo Camera. Figure 4

Illustration of the Euclidean and the polar distance metric

To obtain the vehicle pose, a nonlinear minimization of the error function with respect to the pose parameters t_x , t_z , and θ is performed using the Levenberg–Marquardt algorithm or the Downhill Simplex algorithm [29].

Quaternion-Based Optimization A closed-form solution for the problem of relative orientation between two point clouds with known correspondences is proposed in [30], where quaternions are used for direct parameter estimation. A quaternion $\hat{q}$ is defined by four real numbers q_a , q_b , q_c , and q_d according to

$$\hat{q} = q_a + iq_b + jq_c + kq_d. \quad (5)$$

The last three summands in Eq. 5 can be seen as three “imaginary parts” analogous to the imaginary part of a complex number [31]. Similar to complex numbers, it is $i^2 = j^2 = k^2 = -1$. Unit quaternions are quaternions of norm 1, i.e., $\|\hat{q}\| = 1$. Quaternions allow a much more efficient representation of rotations in 3D space than classical rotation matrices, especially when several rotations are performed subsequently. Calculation rules for quaternions are described in [31].

The measured 3D point cloud is denoted by $\{\vec{w}_{p_i}\}$, the set of corresponding model points by $\{\vec{w}_{m_i}\}$, where $i = 1, \dots, n$ with n as the number of measured 3D points. Subtraction of the centroids yields the zero-mean transformed 3D point clouds $\{\vec{w}'_{p_i}\}$ and $\{\vec{w}'_{m_i}\}$. It is shown in [30] that the determination of the transformation (translation and rotation) between

two point clouds requires the determination of the unit quaternion $\hat{q}$ which maximizes the expression

$$\sum_{i=1}^n (\hat{q} \vec{w}'_{p_i} \hat{q}^*) \cdot \vec{w}'_{m_i} \quad (6)$$

Applying the calculation rules for quaternions to expression (6) yields

$$\hat{q}^T N \hat{q} \quad (7)$$

to be maximized with respect to $\hat{q}$, where

$$N = \sum_{i=1}^n R_{p_i}^T R_{m_i}. \quad (8)$$

In Eq. 8, the matrices R_{p_i} and R_{m_i} are given by

$$R_{p_i} = \begin{pmatrix} 0 & -x'_{p_i} & -y'_{p_i} & -z'_{p_i} \\ x'_{p_i} & 0 & z'_{p_i} & -y'_{p_i} \\ y'_{p_i} & -z'_{p_i} & 0 & x'_{p_i} \\ z'_{p_i} & y'_{p_i} & -x'_{p_i} & 0 \end{pmatrix} \quad (9)$$

$$R_{m_i} = \begin{pmatrix} 0 & -x'_{m_i} & -y'_{m_i} & -z'_{m_i} \\ x'_{m_i} & 0 & z'_{m_i} & -y'_{m_i} \\ y'_{m_i} & -z'_{m_i} & 0 & x'_{m_i} \\ z'_{m_i} & y'_{m_i} & -x'_{m_i} & 0 \end{pmatrix}.$$

The unit quaternion $\hat{q}$ that maximizes expression (8) is then given by the eigenvector $\vec{e}_{\max}$ corresponding to the largest eigenvalue of N . The characteristic polynomial is of degree 4 and can therefore be solved analytically. Based on $\hat{q} = \vec{e}_{\max} = (q_0, q_x, q_y, q_z)^T$, the rotation matrix between the 3D point clouds $\{\vec{w}'_{p_i}\}$ and $\{\vec{w}'_{m_i}\}$ is given by

$$R = \begin{pmatrix} q_0^2 + q_x^2 + q_y^2 - q_z^2 & 2(q_x q_y - q_0 q_z) & 2(q_x q_z + q_0 q_y) \\ 2(q_y q_x + q_0 q_z) & q_0^2 - q_x^2 + q_y^2 - q_z^2 & 2(q_y q_z + q_0 q_x) \\ 2(q_z q_x + q_0 q_y) & 2(q_z q_y + q_0 q_x) & q_0^2 - q_x^2 - q_y^2 + q_z^2 \end{pmatrix}. \quad (10)$$

This analytical approach is computationally highly efficient. For noisy data, however, it suffers from a low robustness as it inherently yields a solution defined by the least-mean-squared Euclidean distance between the points, which is strongly influenced by outliers. Furthermore, when an object model defined by planes is used, the point-to-plane matching can only be performed once before determining the transformation.

As a consequence, it is favorable to combine the analytical optimization technique according to [30] with the ICP approach in order to obtain an algorithm which is faster than the nonlinear optimization scheme but also robust with respect to outliers. The initial pose is again determined based on the centroid and the first principal component of the point cluster representing the moving object. In a first step, point-to-plane matching is performed to obtain the closest point $\tilde{p}_{m_i}$ on the model surface for each 3D point $\tilde{p}_i$, where the Euclidean distance between $\tilde{p}_i$ and $\tilde{p}_{m_i}$ is denoted by $d_i = \|\tilde{p}_i - \tilde{p}_{m_i}\|$. The previously described quaternion-based technique according to [30] yields an analytical solution for the rotation and translation parameters between the 3D point cloud and the cloud of corresponding points on the model surface by minimizing the mean-squared Euclidean distance.

To “imitate” the behavior of an M-estimator in the quaternion-based optimization scheme, each point is weighted with the fair weighting function $w_f(d_i^{(0)})$ prior to the optimization procedure, where $d_i^{(0)}$ is the initial Euclidean distance of the i -th 3D point from the model surface. The fair weighting function is given by

$$w_f(x) = \frac{1}{1 + \frac{|x|}{c}} \quad (11)$$

with c as a user-defined parameter. The corresponding error function E_{quat} then amounts to

$$E_{\text{quat}} = \sum_{i=1}^N \left[\|\tilde{p}_i - \tilde{p}_{m_i}\| w_f(d_i^{(0)}) \right]^2. \quad (12)$$

Once the pose refinement step has been performed, remaining outliers are eliminated from the point cloud.

The quaternion-based optimization procedure is iterated once or twice, where the points are reweighted according to Eq. 12 for each iteration. The behavior of the quaternion-based optimization is similar to that of a direct nonlinear minimization of the Euclidean distances with an M-estimator error function as long as the pose refinement determined by that operation is small.

For the real-world data regarded in section “[Experimental Evaluation](#)” a general observation is that if the original, unweighted 3D points are used in the quaternion-based optimization scheme, the resulting pose accuracy strongly degrades especially for the yaw angle. The computation time of the quaternion-based optimization scheme is about 40 times lower than that of the nonlinear ICP methods (cf. section “[Computational Efficiency](#)”).

Polar Distance Metric (adopted from [36]) The ICP algorithm with the Euclidean metric does not take into account the properties of the stereo measurement process. When reconstructing 3D points from two cameras, the relation between the disparity and the depth is nonlinear. This results in a low accuracy for points which are far away from the camera, while points close to the camera have a higher accuracy. The noise of the point positions is neither Gaussian nor symmetric. The Euclidean metric used in the standard ICP assumes at least a symmetric error distribution of the measurements. To overcome this problem one might think of transforming the geometric model into disparity space (pixel coordinates and disparities), which may be very complex depending on the model. It is thus preferable to use a distance metric which takes into account the nonlinearity of the measurement error.

The 3D points are represented in polar coordinates (cf. Fig. 4b) on the xz plane, where the distance is $r(x, y, z) = \sqrt{x^2 + y^2 + z^2}$ and the polar angle amounts to $\varphi = \arctan(x/z)$. Assuming an ideal pinhole camera model, the pixel coordinates (u, v) can be transformed into the 3D coordinates according to $u/f = x/z$ and $v/f = y/z$, where f denotes the camera constant measured in pixel units. Assuming a standard epipolar geometry, the depth coordinate z can be calculated by $z = b f d$, where b is the baseline distance in meters and d the disparity in pixels. Using these

equations, the distance to the camera $r(u, v, d)$ is given by

$$r(u, v, d) = \sqrt{\left(\frac{bu}{d}\right)^2 + \left(\frac{bv}{d}\right)^2 + \left(\frac{bf}{d}\right)^2}. \quad (13)$$

An error calculation based on the total differential dr is used to analyze the relations of the depth measurement process:

$$dr = \frac{\partial r}{\partial u} du + \frac{\partial r}{\partial v} dv + \frac{\partial r}{\partial d} dd \quad (14)$$

Converting these terms into three-dimensional world coordinate expressions results in the following relations:

$$\begin{aligned} \frac{\partial r}{\partial u} &= \frac{xz}{fr} \\ \frac{\partial r}{\partial v} &= \frac{yz}{fr} \\ \frac{\partial r}{\partial d} &= \frac{z^3}{bfr}. \end{aligned} \quad (15)$$

The term z/r is smaller than or equal to 1. The terms x/f , y/f , and z/f all have the same order of magnitude. The measurement errors du and dv of the pixel coordinates u , v , and the disparity measurement error dd are of the order 0.1–1 pixels and independent of z and r . In our scenario, one can assume that $z/b \gg 1$, whereby the measurement error dr largely depends on the disparity error dd . Accordingly, dr is approximately proportional to z^2 .

Hence, this error analysis yields the result that the normalized measurement error dr/z^2 has an approximately Gaussian distribution. The error distribution of the polar angle φ is also approximately of Gaussian shape. This results in an error function E for the model fit which consists of an error term regarding the distance of a 3D point to the model and a second error term regarding the polar angle difference, combined by the weight factor λ :

$$E(\{r_i\}, \{z_i\}, \{\varphi_i\}) = \sum_{i=1}^N \left[E_r^2(r_i, z_i) + \lambda E_\varphi^2(\varphi_i) \right]. \quad (16)$$

The first error term E_r only depends on distances from the camera, where r_i is the distance of the object point i to the camera and r_{m_i} the distance to the camera

of the intersection point of the same line of sight with the model:

$$E_r(r_i, z_i) = \frac{r_i - r_{m_i}}{z_i^2}. \quad (17)$$

The second error term E_φ only depends on the polar angles:

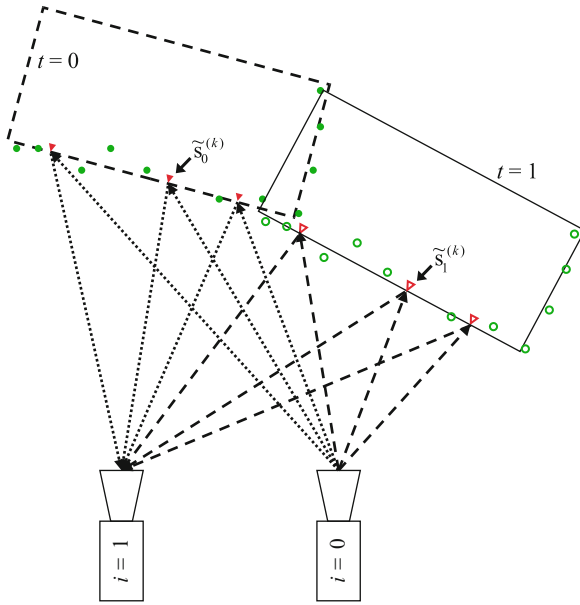
$$E_\varphi(\varphi_i) = \frac{\varphi_i - \varphi_M}{2} (1 + \tanh |\alpha(|\varphi_i - \varphi_M| - \beta)|). \quad (18)$$

The hyperbolic tangent enforces continuity of the angular error function and is required for the convergence behavior of the optimization. The value φ_M denotes the polar angle of the center point of the model. The parameter β corresponds to about half the angular width of the model projected into the image plane, whereas α is a user-defined parameter which influences the optimization behavior. Error function (16) can be minimized in the same manner as the error function involving the Euclidean metric using a nonlinear optimization method.

Computational Efficiency For the ICP approaches described in sections “[Euclidean Distance Metric](#)” and “[Polar Distance Metric](#)” involving standard nonlinear minimization techniques, the repeatedly performed point-to-plane matching and distance evaluation may lead to a significant computational burden of the ICP algorithm with polar coordinates. The C++ implementation requires about 1 ms computation time on a single 2 GHz core for 100 points, where the computation time increases linearly with the number of points. In the application scenario of section “[Experimental Evaluation](#),” the object is typically represented by 200–400 points. The computational efficiency of the quaternion-based optimization scheme (cf. section “[Quaternion-Based Optimization](#)”) is about 40 times higher than that of the nonlinear ICP methods.

Model-Based Scene Flow

In this section, a temporal extension of the concept of model-based stereo is described, which is termed model-based scene flow. This method aims for a computation of the 3D pose of the object and an instantaneous estimation of its temporal derivative using two subsequent stereo image pairs, thus avoiding



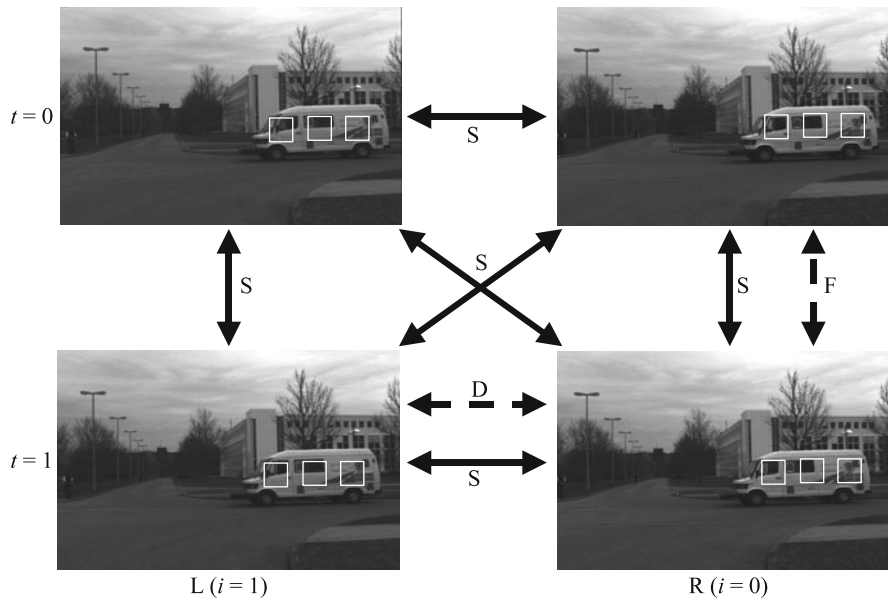
3D Pose Estimation of Vehicles Using Stereo Camera. Figure 5

Principle of model-based scene flow. *Circles*: 3D points, *triangles*: corresponding points on the model surface

a temporal filtering stage. To initialize the lateral velocity $\dot{t}_x$ of the vehicle, information provided by the optical flow field of the right camera is used. The velocity $\dot{t}_z$ along the depth axis and the yaw rate $\dot{\theta}$ are initially unknown, such that their values are initialized by zero.

The basic idea of model-based scene flow estimation is to use the model surface together with the pose parameters and their temporal derivatives to obtain the homographies that determine image points corresponding to each other in all four images. Pose and motion parameters are used to connect images taken at the same time by different cameras and images acquired at different times, respectively (cf. Fig. 5). Image windows of predefined size around the K reprojected points are extracted from the four images (cf. Fig. 6). Each of the four images I_{it} is identified by two indices i and t , where the right image is denoted by the index $i = 0$ and the left image by $i = 1$, while the previous time step is denoted by the index $t = 0$ and the current time step by $t = 1$.

The 3D model is assumed to be composed of planes, which is always the case for tessellated models



3D Pose Estimation of Vehicles Using Stereo Camera. Figure 6
Determination of optical flow (F), disparity (D), and scene flow (S)

like the cuboid used in section “[Experimental Evaluation](#).” The projective vector $\tilde{q}_{i't'}^{(k)}$ of the pixel coordinates of window k in image i' at time step t' is obtained from the corresponding projective vector $\tilde{q}_{it}^{(k)}$ in image i at time step t by the homography ${}_{it}^{i't'}H^{(k)}$ induced by the model plane on which the center of window k is located. It is thus

$$\tilde{q}_{i't'}^{(k)} = {}_{it}^{i't'}H^{(k)} \tilde{q}_{it}^{(k)} \quad (19)$$

with $k = 1, \dots, K$. The homography ${}_{it}^{i't'}H^{(k)}$ depends on the known extrinsic and intrinsic camera parameters and on the pose at time steps t and t' of the model plane on which patch k is located, which follows from the 3D pose parameters of the object and their temporal derivatives.

A detailed description of the concept of the homography induced by a plane is given in [32]. Effectively, the determination of the homography ${}_{it}^{i't'}H^{(k)}$ between the views of camera i at time step t and camera i' at time step t' relies on the following procedure (cf. Fig. 5): The first step is the computation of the intersection point $\tilde{s}_t^{(k)}$, at time step t between the ray that belongs to the image point $\tilde{q}_{it}^{(k)}$ and the model plane at time step t on which the center of window k is located. If $t \neq t'$, the point $\tilde{s}_t^{(k)}$, which is a 3D point in the scene, has to be adjusted according to the object motion between the time steps t and t' , resulting in the point $\tilde{s}_{t'}^{(k)}$. This adjustment consists of a rotation and a translation readily inferred from the temporal derivatives of the pose parameters. If $t = t'$, $\tilde{s}_{t'}^{(k)} = \tilde{s}_t^{(k)}$. The point $\tilde{s}_{t'}^{(k)}$ is then projected into camera i' , resulting in the corresponding image point $\tilde{q}_{i't'}^{(k)}$.

The error function E_{msf} minimized by the model-based scene flow algorithm consists of the sum over the six pairwise similarity measures between the projected image patches according to

$$E_{\text{msf}} = \sum_{k,i,t,t' \neq i,t'} S\left(I_{it}\left(\tilde{q}_{it}^{(k)}\right), I_{i't'}\left(\tilde{q}_{i't'}^{(k)}\right)\right) \quad (20)$$

where S denotes the similarity measure, $I_{it}\left(\tilde{q}_{it}^{(k)}\right)$ the window around the point $\tilde{q}_{it}^{(k)}$ in image i at time step t , and $I_{i't'}\left(\tilde{q}_{i't'}^{(k)}\right)$ the corresponding window in image i' at time step t' , warped by the homography ${}_{it}^{i't'}H^{(k)}$. Warping is required to account for the difference in perspective between the views.

The zero-mean sum of squared differences (SSD) and the normalized cross-correlation coefficient (NCC) are used as the similarity measure S . A Sobel edge detector is used as an interest operator to select windows which display a significant amount of contrast. The parameters describing the 3D pose and its temporal derivatives are obtained by minimizing (SSD) or maximizing (NCC) Eq. (20) using the Gauss–Newton algorithm. If only one stereo image pair is regarded, optimization of Eq. 20 yields a pose estimation according to the concept of model-based stereo.

The computational efficiency of the model-based scene flow is typically comparable to that of the ICP approach with classical optimization (i.e., without quaternion) as described in sections “[Euclidean Distance Metric](#)” and “[Polar Distance Metric](#).”

Experimental Evaluation

This section presents an evaluation of the ICP-based approach and the model-based scene flow stage. It partially draws upon the description in [36]. Three color cameras with a resolution of $1,034 \times 776$ pixels were used for image acquisition. The cameras were mounted side by side at different displacements, resulting in three different baseline distances b of 0.102, 0.228, and 0.380 m. The frame rate of the sequences is about 14 fps. The color information is only used for ground truth calculation, whereas for the stereo algorithms the images were converted to grayscale. Several colored markers have been attached to the vehicle, and a color classifier was used to extract the markers in the images. Afterward, the corner positions in each image were estimated by a highly accurate corner detector [33] at sub-pixel accuracy. The corner positions were used for bundle adjustment to compute the three-dimensional world coordinates of each corner point. Four markers were attached to each side of the vehicle, such that the corresponding plane is overdetermined. The colored markers hardly display a contrast in the grayscale images and therefore do not generate additional 3D points or optical flow vectors.

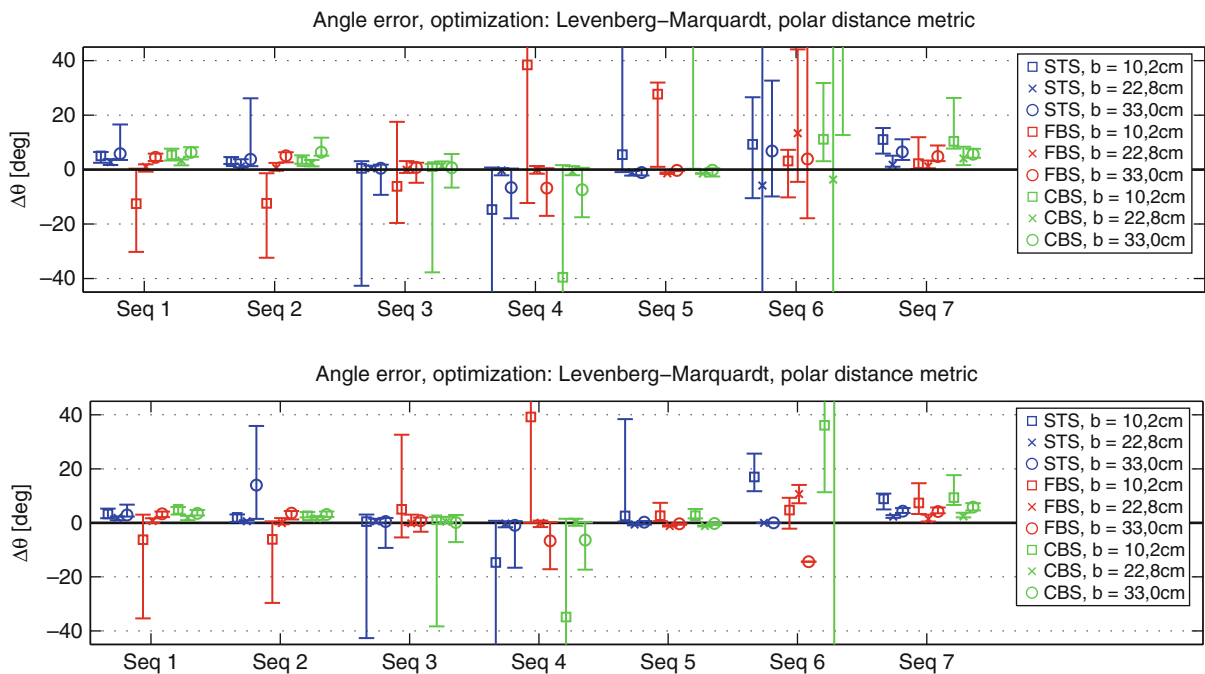
The system performs a frame-by-frame 3D pose estimation without any temporal filtering (tracking by detection). The 3D point cloud is clustered to obtain

the scene part which represents the vehicle. The first principal component of this point cloud is used to initialize the angular pose parameter.

A set of seven different scenes, each of which was recorded with three different stereo baseline distances, is used. These sequences display a moving vehicle on a road crossing in typical intersection situations, either passing straight through the intersection or turning left or right at different velocities, where the distance of the object from the camera amounts to 15–30 m. (The images and reference data are publicly available at <http://aiweb.techfak.uni-bielefeld.de/content/vehicle-pose-data-set>.) The evaluation is based on two geometric indicators: the yaw angle difference between the true yaw angle and the pose estimation result and the mean distance of the ground truth points to the corresponding model planes. The differences are visualized using

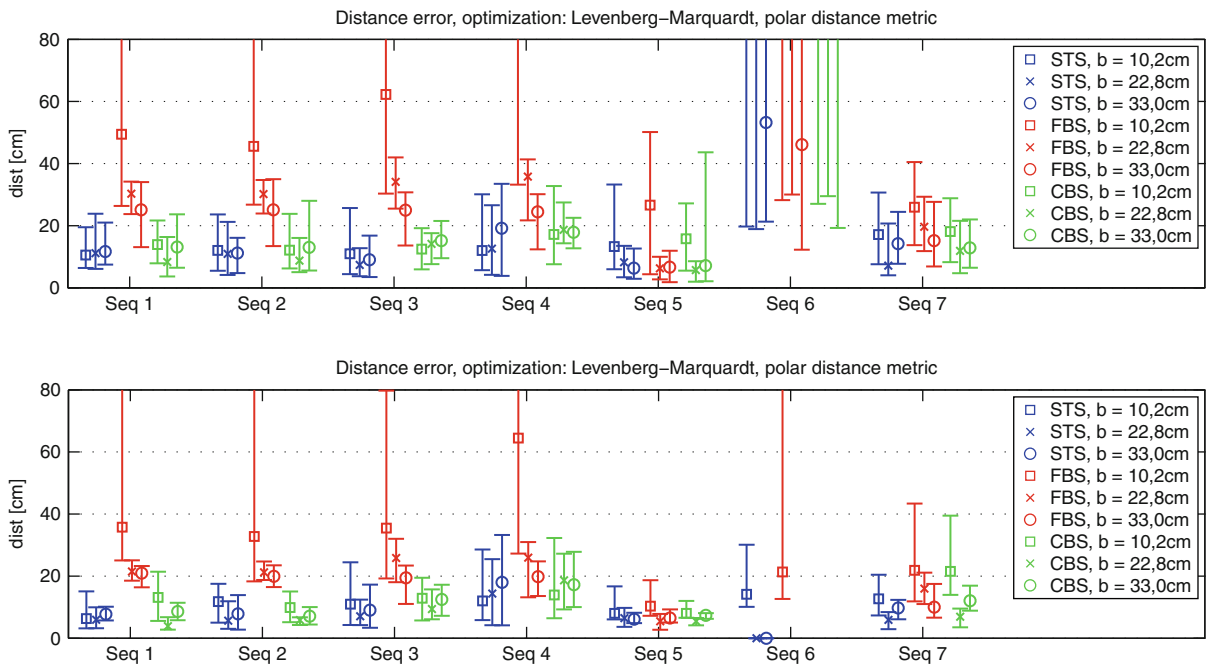
error bars denoting the median of the full sequence and the 25% and the 75% quantiles, respectively.

For the images rectified to standard stereo geometry, the effective principal distance amounts to $f = 1,350$ pixels, such that $bf = 307.8$ m pixel for the medium baseline distance of $b = 0.228$ m. According to [33], the difference between two checkerboard corner positions (i.e., a disparity value) can be measured with an accuracy (here corresponding to one-half of the difference between the 25% and the 75% quantiles) of $\Delta d = 0.032$ pixels. At a typical object distance from the camera of $z_0 = 20$ m, the disparity thus amounts to $d_0 = bf/z_0 = 15.39$ pixels, resulting in a depth difference accuracy of $\sqrt{2}[bf/d_0 - bf/(d_0 + \Delta d)] = 0.059$ m (cf. law of error propagation). The distance in the scene between the pairs of markers attached to the frontal and the rear part of the vehicle amounts to about 3 m, such that the



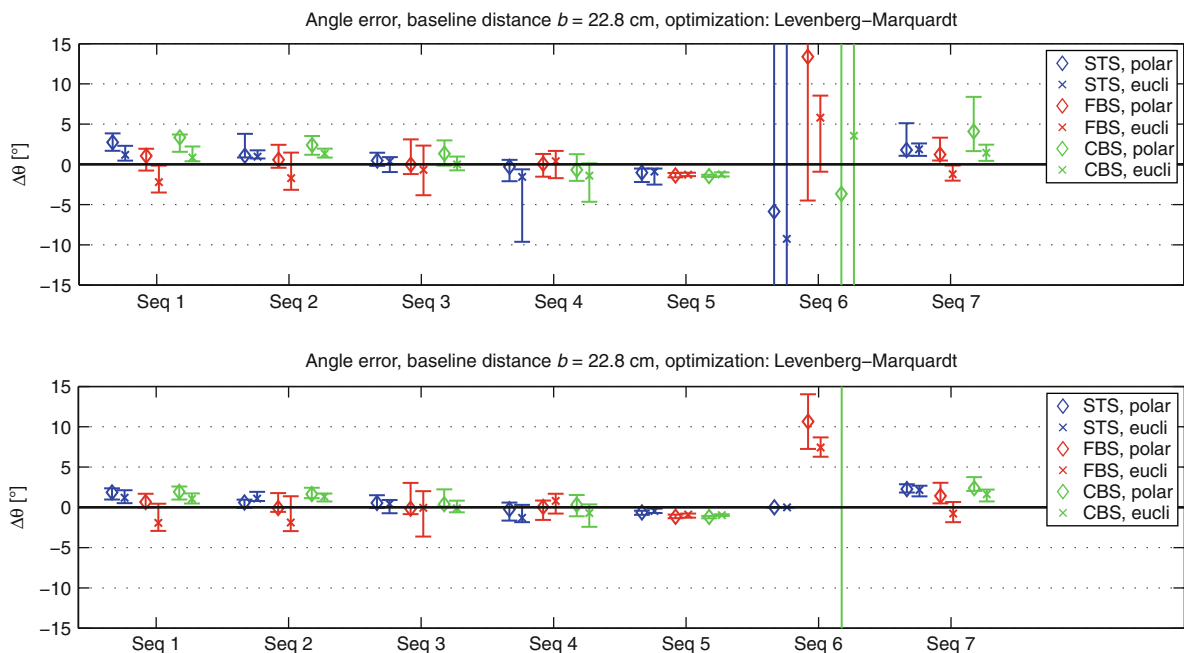
3D Pose Estimation of Vehicles Using Stereo Camera. Figure 7

Dependence of the yaw angle error (denoted by $\Delta\theta$) on the stereo baseline; Levenberg–Marquardt optimization, polar distance metric. *Upper diagram*: all images; *lower diagram*: rear views excluded. Blue color denotes the spacetime stereo technique (STS), red color the feature-based stereo (FBS), and green color the correlation-based stereo (CBS). The small baseline is marked by squares, the intermediate baseline by crosses, and the large baseline by circles. The intermediate baseline distance is most favorable with respect to the yaw angle error



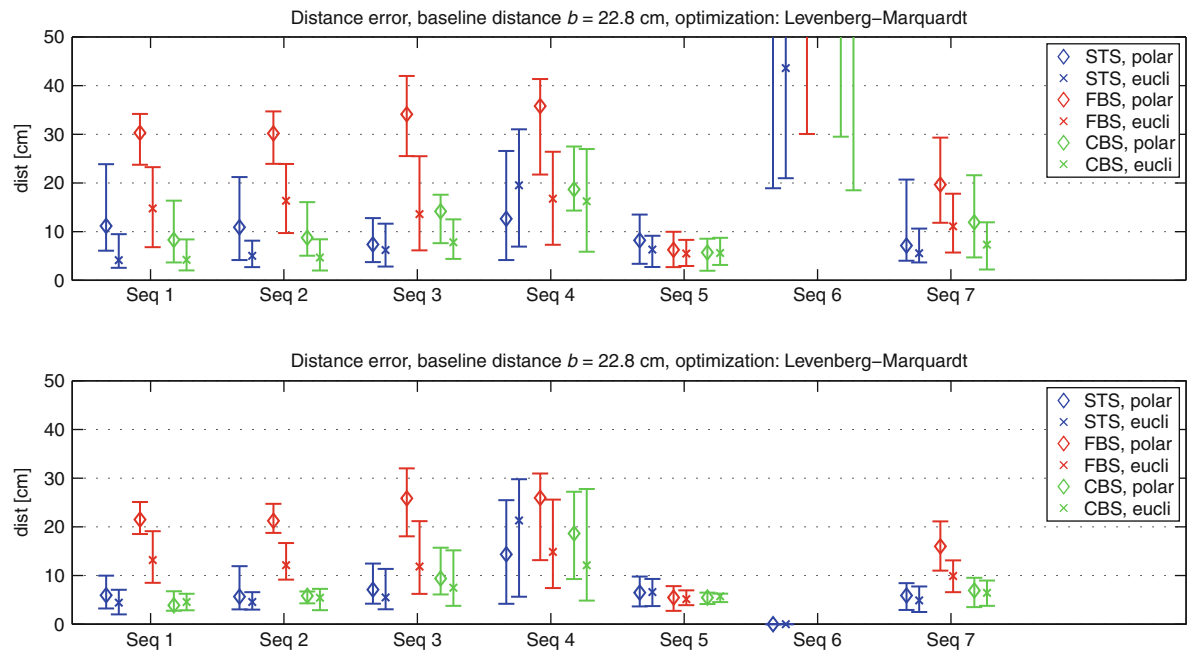
3D Pose Estimation of Vehicles Using Stereo Camera. Figure 8

Dependence of the distance error on the stereo baseline; Levenberg–Marquardt optimization, polar distance metric. *Upper diagram:* all images; *lower diagram:* rear views excluded. Colors and markers as in Fig. 7



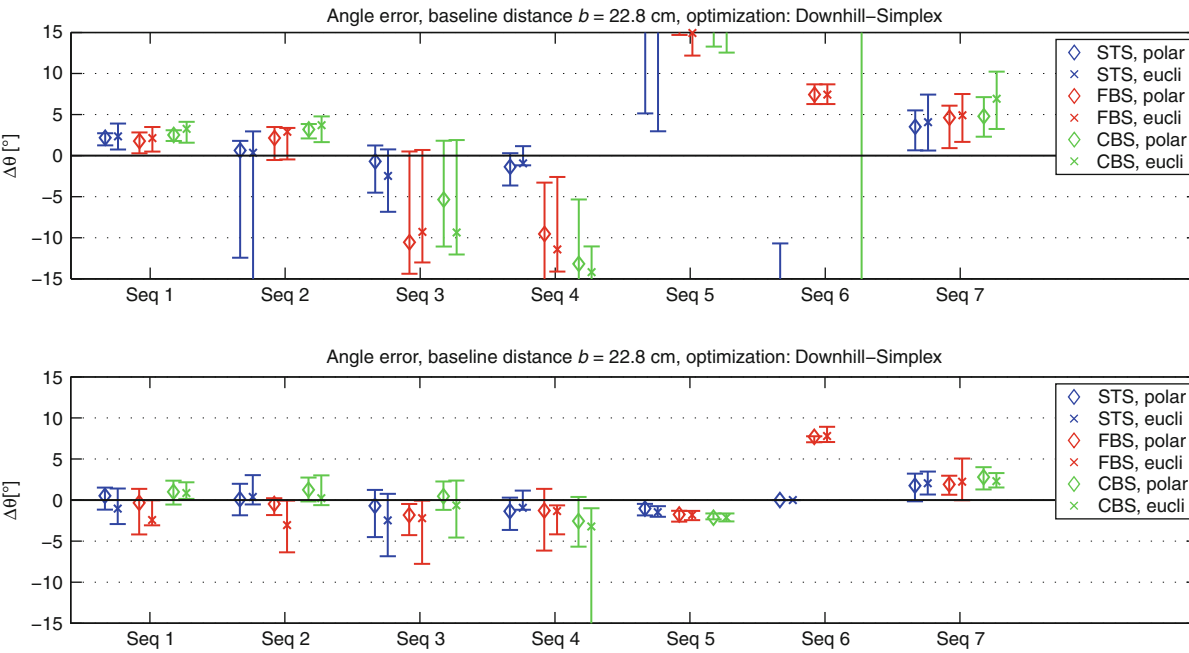
3D Pose Estimation of Vehicles Using Stereo Camera. Figure 9

Dependence of the yaw angle error (denoted by $\Delta\theta$) on the stereo algorithm; Levenberg–Marquardt optimization, intermediate stereo baseline. *Upper diagram:* all images; *lower diagram:* rear views excluded. The colors (blue, red, and green) denote the stereo algorithm, while diamonds represent the polar metric and stars the Euclidean metric



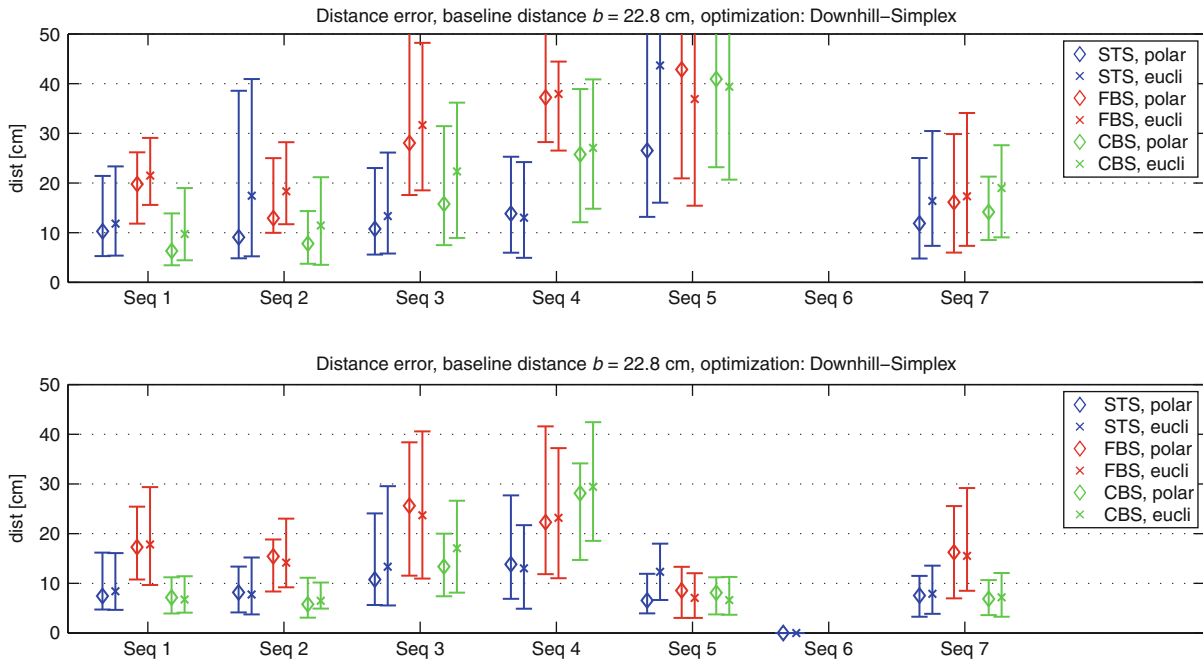
3D Pose Estimation of Vehicles Using Stereo Camera. Figure 10

Dependence of the distance error on the stereo algorithm; Levenberg–Marquardt optimization, intermediate stereo baseline. *Upper diagram*: all images; *lower diagram*: rear views excluded. Colors and markers as in Fig. 9



3D Pose Estimation of Vehicles Using Stereo Camera. Figure 11

Yaw angle error for Downhill Simplex optimization, intermediate stereo baseline. *Upper diagram*: all images; *lower diagram*: rear views excluded. Colors and markers as in Fig. 9



3D Pose Estimation of Vehicles Using Stereo Camera. Figure 12

Distance error for Downhill Simplex optimization, intermediate stereo baseline. *Upper diagram: all images; lower diagram: rear views excluded.* Colors and markers as in Fig. 9

lowest accuracy of the yaw angle, which occurs when the vehicle is observed laterally, corresponds to 1.1° . The absolute depth and yaw angle errors are proportional to z_0^2 .

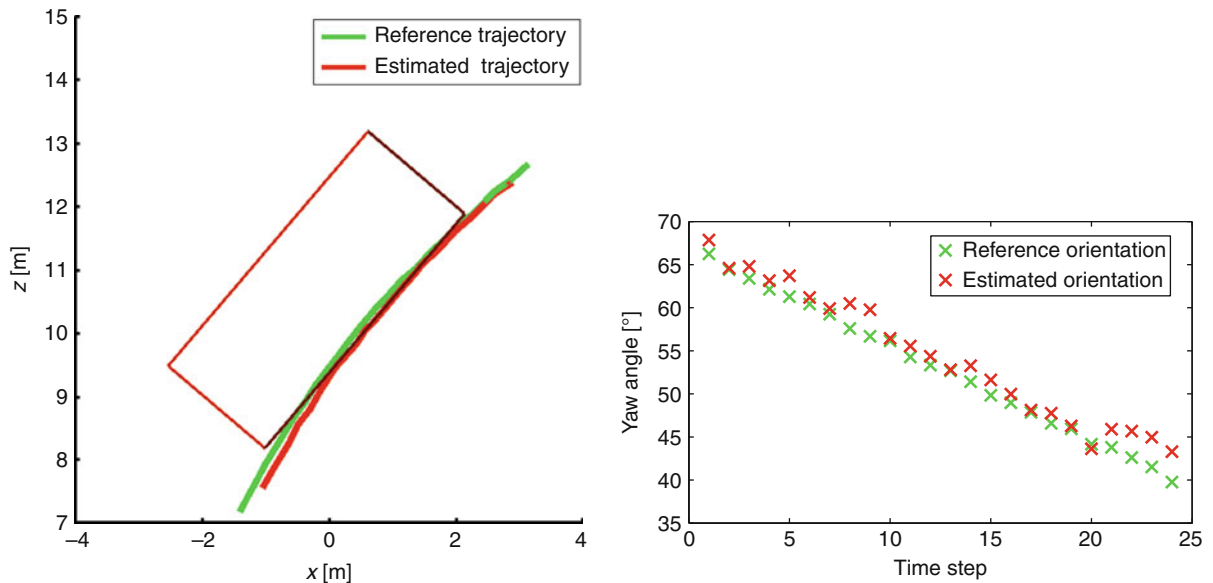
Evaluation of the ICP-Based 3D Pose Estimation

This section provides an evaluation of the ICP-based approach with respect to the influence of the baseline distance of the stereo camera system, the employed stereo algorithm, the distance metric, and the optimization algorithm used for minimizing the error function with respect to the pose parameters. In Figs. 7–12, the median values and 25% and 75% quantiles of the deviations of the inferred pose parameters from their ground truth values are determined (1) based on all images of each sequence (upper diagram) and (2) based on all images in which the extension of the 3D point cloud in x direction is at least 20% larger than the vehicle width (lower diagram). The second evaluation has been added because the pose estimation result is expected to be rather

unreliable for frontal and rear views of the vehicle due to the resulting small number of 3D points on its longer side.

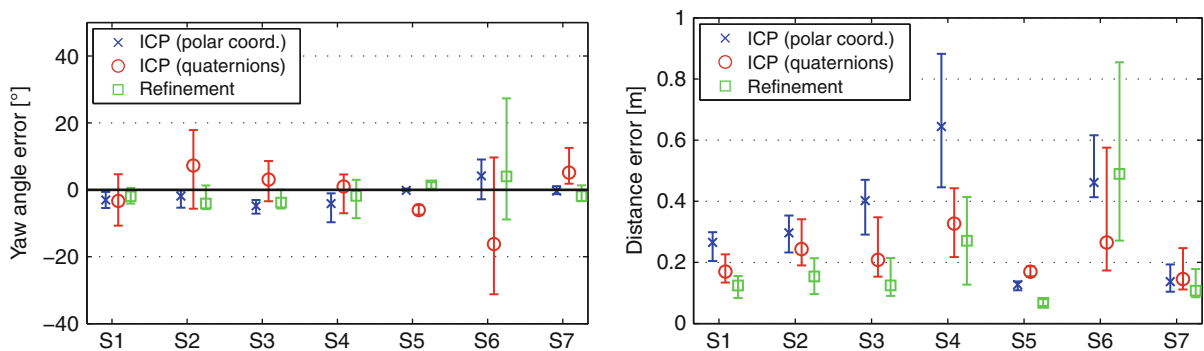
Baseline Distance (adopted from [36]) The first part of the evaluation analyzes the different baseline distances of the sequences. Figure 7 shows the yaw angle error and Fig. 8 the distance error for the three different baseline distances for all seven sequences using the three different stereo approaches described. The comparison between the different baseline distances shows that the smallest baseline is not sufficient for our application. Especially for vehicles turning left or right in front of the camera (sequences 3 and 4) the errors are large. The largest baseline also produces larger errors, since due to the large camera displacement and the resulting parallax effects the stereo algorithms produce a smaller number of correct correspondences.

Stereo Algorithm (adopted from [36]) To evaluate the different stereo algorithms, tracking results are



3D Pose Estimation of Vehicles Using Stereo Camera. Figure 13

Left: Trajectory for sequence S5 with the vehicle model drawn for one time step, depicting the position of the center of the lateral model surface. *Right:* Behavior of the yaw angle for the same sequence



3D Pose Estimation of Vehicles Using Stereo Camera. Figure 14

Yaw angle error (*left*) and distance error (*right*). Symbols denote the median values and error bars the 25% and 75% quantiles. "Refinement" denotes the result of model-based scene flow

computed for all sequences with the intermediate base-line distance and the Levenberg–Marquardt algorithm. Figures 9 and 10 show the results of the different stereo algorithms, where both distance metrics (Euclidean and polar) are evaluated. The results for the yaw angle error show that the difference between the pixel accurate feature-based stereo and the other two algorithms with sub-pixel accuracy is negligible. The results are similar for all three stereo algorithms. However,

regarding the vehicle position, a difference between pixel and sub-pixel accuracy is noticeable. The distances of the feature-based stereo results show an up to three times higher error for sequences 1, 2, and 3. No significant difference between the spacetime stereo and the correlation-based stereo can be found.

Distance Metric (adopted from [36]) Figures 9 and 10 show the pose estimation errors for the two distance

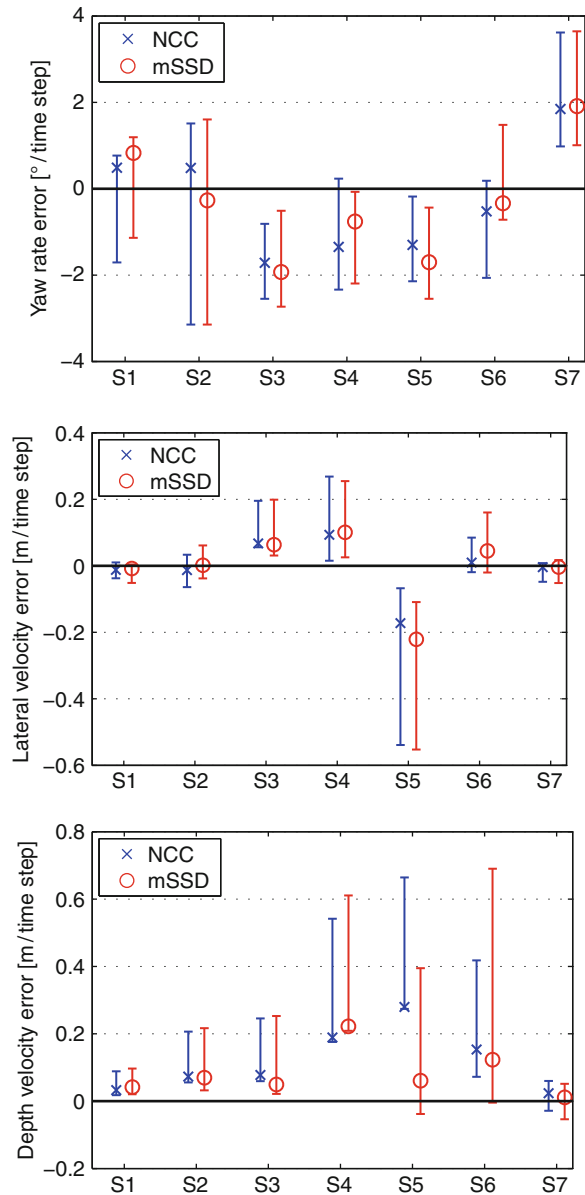
metrics. For the position error both metrics have a similar performance, whereas the polar distance metric produces a smaller yaw angle error for most sequences. The computation time for the polar metric is the same as for the Euclidean metric.

Optimization Algorithm (adopted from [36]) The results of the Downhill Simplex optimization are shown in Fig. 11 for the yaw angle error and in Fig. 12 for the error of the average distance between the ground truth points and the corresponding model planes. The yaw angle errors are all higher for the Downhill Simplex algorithm. Considering that the number of function calls of the Levenberg–Marquardt algorithm is of the same order of magnitude (about 100 function calls) as for the Downhill Simplex algorithm with a better pose estimation accuracy, the Levenberg–Marquardt algorithm appears to be preferable. The distance errors show a similar behavior.

Evaluation of the Model-Based Scene Flow Stage

For the evaluation of the model-based scene flow stage, an initial pose is computed using the quaternion-based ICP algorithm according to section “Quaternion-Based Optimization.” The stereo baseline is set to $b = 0.228$ m, corresponding to the configuration which turned out to yield the most accurate pose estimation results according to section “Evaluation of the ICP-Based 3D Pose Estimation.” Figure 13 depicts an example trajectory and the estimated yaw angle for sequence S5. The reference trajectory is obtained using the average position of the four colored markers. The estimated trajectory is generated based on the center of the corresponding model surface.

The reference data for the temporal pose derivatives are obtained by numerical differentiation of the pose reference data with respect to time. Since the resulting yaw rate is too noisy to provide sufficiently reliable reference values, a fourth-order polynomial fitted to the differentiated yaw angle reference values is used as reference data for the yaw rate. In Figs. 14 and 15, the deviations between the estimated pose parameters and their temporal derivatives and the corresponding reference values are visualized by error bars denoting the median over the sequence and the 25% and the 75% quantiles, respectively. The model-based scene flow



3D Pose Estimation of Vehicles Using Stereo Camera.

Figure 15

Error of the yaw rate (top), horizontal velocity (middle), and the velocity along the depth axis (bottom). Symbols denote the median values and error bars the 25% and 75% quantiles

method is always initialized with the result of the quaternion-based ICP approach.

The average yaw angle accuracy of the nonlinear ICP approach with polar coordinates is about three



3D Pose Estimation of Vehicles Using Stereo Camera. Figure 16
Example sequence showing a passenger car turning right

times higher than that of the quaternion-based solution, i.e., $2.4^\circ \pm 2.1^\circ$ versus $4.3^\circ \pm 6.3^\circ$ (cf. Fig. 14, top). The average yaw angle error of the model-based scene flow method amounts to $1.8^\circ \pm 3.4^\circ$. The fairly large error interval is mainly caused by the high uncertainty observed for sequence S6 while the errors are comparable to those of the nonlinear ICP approach for all other sequences. The average positional error of the quaternion-based ICP is lower by one third than that of the nonlinear ICP (0.21 ± 0.07 m vs. 0.31 ± 0.08 m); the model-based scene flow method yields a comparable average error value of 0.19 ± 0.11 m (cf. median values in Fig. 14, bottom).

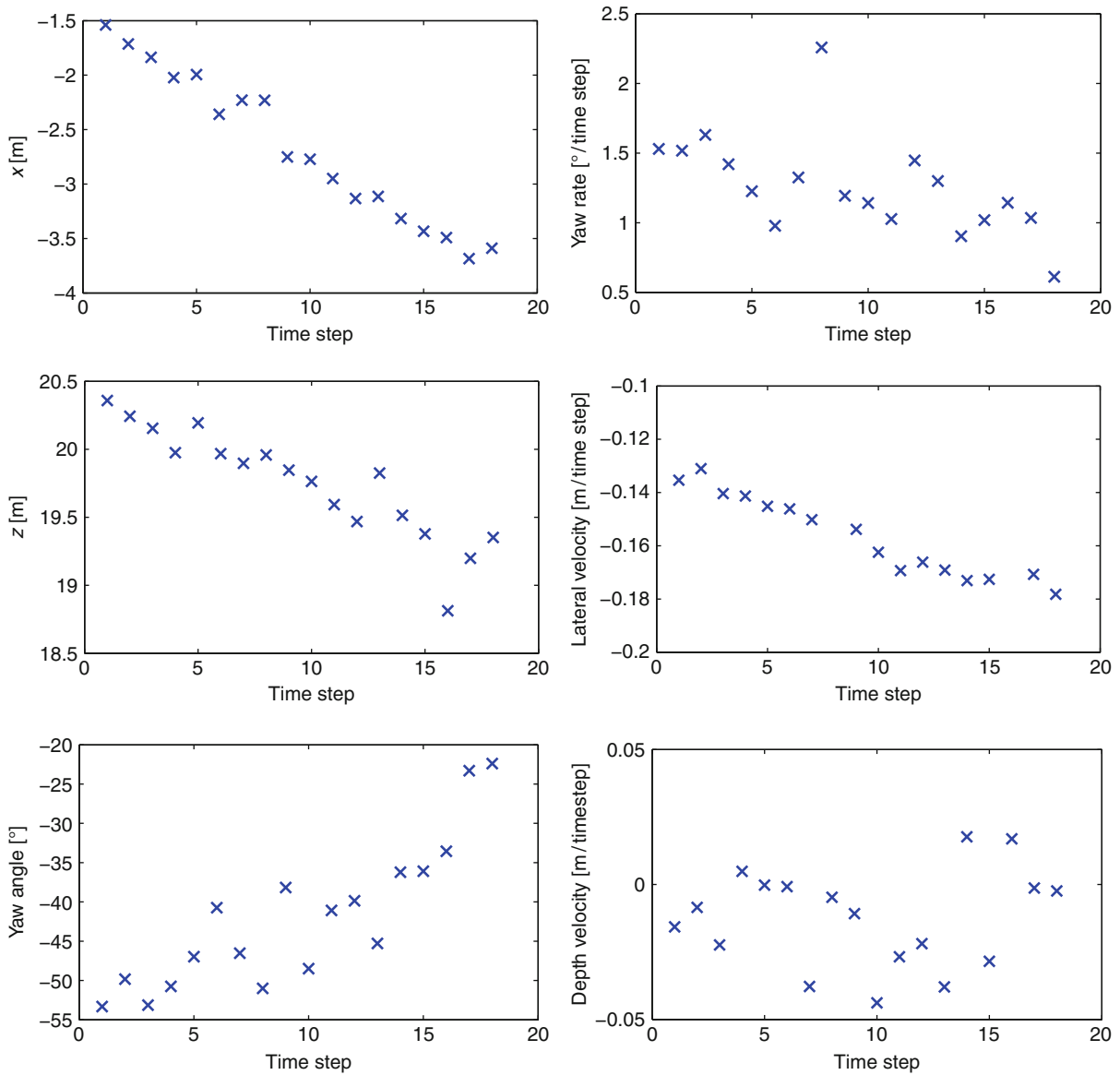
The accuracy of the yaw rate estimated with the model-based scene flow approach, averaged over all test sequences, corresponds to $1.2^\circ \pm 1.3^\circ$ per time step (cf. Fig. 15). The average horizontal velocity error amounts to 0.06 ± 0.07 m per time step, while the average error of the velocity component along the depth axis is somewhat higher and corresponds to 0.11 ± 0.09 m per time step. For the model-based scene flow method, the evaluation results do not show significant differences between the zero-mean SSD and the NCC similarity measures. Comparably large errors mainly occur for sequences S4, S5, and S6, where in most images the object is far away from the camera ($z \approx 30$ m) and only its rear side is visible.

As a further example, a short image sequence of 1.5 s length is regarded, showing a passenger car performing a turning maneuver (cf. Fig. 16). The image size amounts to 640×480 pixels, the frame rate is 12 fps, and the baseline distance corresponds to 0.3 m. In the regarded traffic setting, no photogrammetric markers could be attached to the vehicle and

thus no reference data are available. The position (t_x, t_z) and the yaw angle θ determined with the model-based scene flow method illustrate that the vehicle moves laterally from the right to the left, approaches the camera, and performs a rotation by about 25° , corresponding to an average yaw rate of about 1.4° per time step (cf. Fig. 17). The measurements of the instantaneous yaw rate $\dot{\theta}$ are somewhat noisy but consistently indicate a value slightly decreasing from about 1.5 to 1.0° per time step. The measured instantaneous lateral velocity $\dot{t}_x$ of 0.14 – 0.18 m per time step is consistent with the behavior of the lateral position t_x over time. The measurements of the depth velocity $\dot{t}_z$ are rather noisy but illustrate that the model-based scene flow method is able to instantaneously detect small velocity components along the depth axis of only a few centimeters per time step, corresponding to less than a percent of the depth itself. At the end of the sequence, the depth velocities are close to zero since the vehicle moves more laterally with respect to the camera than at the beginning.

Summary and Conclusion

This study has described and evaluated approaches for three-dimensional pose estimation of vehicles using a stereo camera. After computation of stereo and optical flow information, a graph-based clustering stage in the space spanned by the spatial and velocity coordinates separates the static from the moving objects in the scene. An iterative closest point algorithm (ICP) estimates the vehicle pose using a cuboid as a weak vehicle model. A classical nonlinear optimization and a computationally efficient quaternion-based analytical



3D Pose Estimation of Vehicles Using Stereo Camera. Figure 17

Pose parameters and their temporal derivatives estimated with the model-based scene flow method for the image sequence shown in Fig. 16

approach relying on the Euclidean distance metric have been compared to a distance metric based on polar coordinates which takes into account the properties of the stereo measurement process. The tracking-by-detection approach has been employed, such that no temporal filtering (e.g., Kalman filtering) has been applied. The algorithm has been evaluated on seven

different real-world sequences with respect to different stereo algorithms, baselines, distance metrics, and optimization algorithms. The first part of the evaluation has regarded all images of all sequences, while in a second part the images in which the vehicle is oriented longitudinally with respect to the camera have been excluded. For most of the sequences, the pose

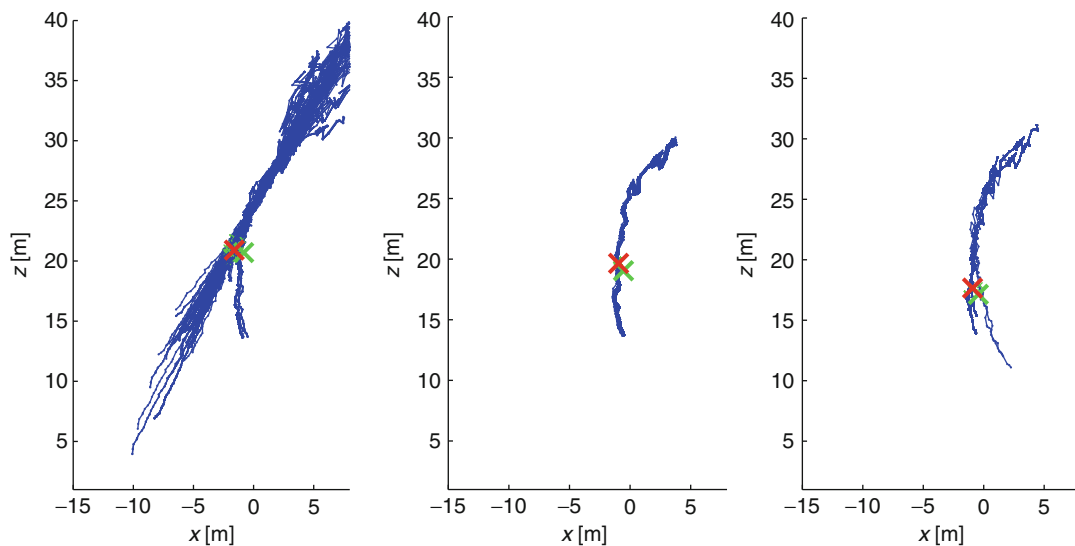
estimation yields reliable results for about 90% of the time steps. Only in sequence S6 the vehicle drives nearly parallel to the z axis, leading to a small projected extension of the 3D point cloud in x direction, resulting in unreliable pose estimation results for more than 90% of the frames of that sequence.

Furthermore, a method for model-based estimation of 3D pose parameters and their instantaneous temporal derivatives has been examined in the context of moving vehicle detection. The quaternion-based ICP algorithm has been used for initialization. Pose refinement and estimation of the temporal pose derivatives are again performed without relying on temporal filtering, using a model-based scene flow method where 3D points on the object surface are reprojected into pairs of stereo images acquired at subsequent time steps and similarities between image windows extracted around the reprojected points are maximized.

The evaluation shows that for obstacles in a range between 10 and 30 m ahead of the ego-vehicle, the intermediate baseline distance of 0.228 m is preferable. In that scenario a sub-pixel accurate stereo algorithm yields up to three times higher distance accuracies when compared to a pixel accurate algorithm. To

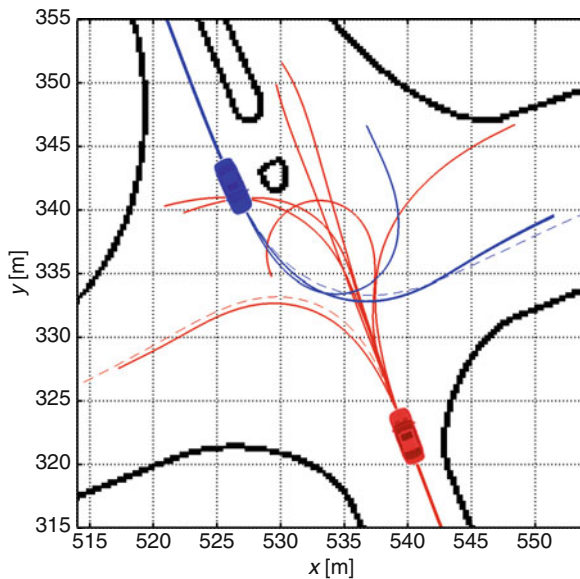
minimize the point distances in the ICP algorithm, the Levenberg–Marquardt algorithm yields better accuracies than the Downhill Simplex optimization while the number of function calls is comparable. The proposed polar distance metric is preferable to the Euclidean distance metric as for most sequences it yields a higher accuracy of the estimated yaw angle, especially for the pixel accurate stereo algorithm. If this most favorable configuration is employed, the pose estimation yields yaw angle errors typically smaller than 3° and absolute distance errors of about 0.1 m, corresponding to relative distance errors around 0.5%. For the model-based scene flow method, the yaw angle and positional errors are similar. The accuracy of the estimated instantaneous yaw rate is about 1° per time step, while the accuracies of the lateral velocity and the velocity along the depth axis are both of the order of 0.1 m per time step.

These evaluation results indicate that the described methods for image-based 3D pose estimation of vehicles yield useful information for driver assistance systems in which instantaneous knowledge about the position and motion direction of vehicles is of essential importance.



3D Pose Estimation of Vehicles Using Stereo Camera. Figure 18

Particles (blue) for a vehicle driving across a roundabout (time $t = 0.0$ s, 0.1 s, 0.3 s, left to right) as obtained with the particle filter method described in [34]. The estimated vehicle position (red) is compared with the true position (green) (from [34])



3D Pose Estimation of Vehicles Using Stereo Camera.
Figure 19

Example of an early stage of a situation in which both vehicles are turning left (cf. [35]). The future positions (ground truth) of the vehicles are represented by the dashed lines, while the solid lines represent the predictions (from [35])

Future Directions

In driver assistance systems for complex traffic scenarios such as road intersections or roundabouts in the urban environment, highly dynamical motion patterns need to be predicted several seconds into the future. Such predictions may serve for the recognition of driving maneuvers or the analysis of situations involving several vehicles with respect to the possible occurrence of collisions.

As an example of ongoing work in this field, the object position is predicted in [34] based on a Bayesian framework in which the probability distribution of the motion hypotheses is represented by a set of measured example trajectories (particles) rather than motion models. These particles are propagated over time using a particle filter [18, 34]. The measurement noise is inherent in the example trajectories and therefore does not need to be modeled. An example result is shown in Fig. 18. Effectively, this method avoids using

simple motion models by cognitive prediction of motion patterns based on a large data base of known motion behaviors.

A further step is the recognition and prediction of situations involving two or more vehicles based on their 3D poses simultaneously acquired over time. Recent work [35] focuses on the recognition of situations involving two vehicles at road intersections. For each vehicle, a set of possible future motion trajectories is predicted several seconds ahead by the trajectory particle filter using a motion database (cf. Fig. 19). An interaction model based on the mutual visibility of the vehicles and the assumption that the drivers will attempt to avoid a collision is used to assign probabilities to possible future situations (cf. Fig. 19), allowing a classification of the situation by a probabilistic framework.

These short descriptions of research activities in the fields of cognitive motion prediction, maneuver recognition, and situation classification illustrate possible future directions of the evolution of driver assistance systems. Fundamentally relying on 3D pose estimation results, such systems will anticipate the behavior of the traffic participants and predict their positions, motion patterns, individual intentions, and mutual interactions.

Bibliography

1. Collado JM, Hilario C, de la Escalera A, Armingol JM (2004) Model based vehicle detection for intelligent vehicles. In: Proceedings of the IEEE intelligent vehicles symposium, Parma, pp 572–577
2. Dellaert F, Thorpe CE (1997) Robust car tracking using kalman filtering and Bayesian templates. In: Proceedings of the conference on intelligent transportation systems, Boston
3. Koller D, Danilidis K, Nagel H-H (1993) Model-based object tracking in monocular image sequences of road traffic scenes. *Int J Comput Vision* 10(3):257–281
4. Kollnig H, Nagel H-H (1997) 3D pose estimation by directly matching polyhedral models to gray value gradients. *Int J Comput Vision* 23(3):283–302
5. Haag M, Nagel H-H (1999) Combination of edge element and optical flow estimates for 3D model-based vehicle tracking in traffic image sequences. *Int J Comput Vis* 35(3):295–319
6. Kato T, Ninomiya Y, Masaki I (2002) An obstacle detection method by fusion of radar and motion stereo. *IEEE Trans Intell Transp Syst* 3(3):182–188

7. Steux B, Lurgeau C, Salesse L, Wautier D (2002) Fade: a vehicle detection and tracking system featuring monocular color vision and radar data fusion. *Proceedings of the IEEE Intell Vehicle Symp* 2:632–639
8. Wender S, Dietmayer K (2007) 3D vehicle detection using a laser scanner and a video camera. In: *Proceedings of the sixth European congress on ITS in Europe*, Aalborg
9. Kaempchen N (2007) Feature-level fusion of laser scanner and video data for advanced driver assistance systems. PhD thesis, Faculty of Engineering and Computer Science, Ulm University
10. Knöppel C, Regensburger U, Schanz A, Michaelis B (2000) Robuste Erkennung von Straßenfahrzeugen im Rückraumbereich eines Straßenfahrzeuges. In: *Proceedings of the 22nd DAGM symposium*. Springer, Heidelberg, pp 35–42
11. Dang T, Hoffmann C, Stiller C (2002) Fusing optical flow and stereo disparity for object tracking. In: *Proceedings of the IEEE international conference on intelligent transportation systems*, Singapore
12. Leibe B, Cornelis N, Cornelis K, van Gool L (2006) Integrating recognition and reconstruction for cognitive traffic scene analysis from a moving vehicle. In: *Proceedings of the 28th DAGM annual pattern recognition symposium*. Springer, Heidelberg, pp 192–201
13. Leibe B, Schiele B (2004) Scale-invariant object categorization using a scale-adaptive mean-shift search. In: *Proceedings of the 26th DAGM annual pattern recognition symposium*. Springer, Heidelberg, pp 145–153
14. Toulminet G, Bertozzi M, Mousset S, Bensrhair A, Broggi A (2006) Vehicle detection by means of stereo vision-based obstacles features extraction and monocular pattern analysis. *IEEE Trans Image Process* 15(8):2364–2375
15. Kubota S, Nakano T, Okamoto Y (2007) A global optimization algorithm for realtime on-board stereo obstacle detection systems. In: *Proceedings of the IEEE intelligent vehicles symposium*, Istanbul, pp 7–12
16. Nedeveschi S, Danescu R, Marita T, Oniga F, Pocol C, Sobol S, Tomiuc C, Vancea C, Meinecke MM, Graf T, To TB, Obojski MA (2007) A sensor for urban driving assistance systems based on dense stereovision. In: *Proceedings of the IEEE intelligent vehicles symposium*, Istanbul, pp 276–283
17. Barth A, Franke U (2008) Where will the oncoming vehicle be the next second? In: *Proceedings of the IEEE intelligent vehicles symposium*, Eindhoven
18. Hermes C, Barth A, Wöhler C, Kummert F (2009) Object motion analysis and prediction in stereo image sequences. In: *Proceedings of the Oldenburger 3D-Tage*, Oldenburg
19. Barth A, Siegemund J, Franke U, Förstner W (2009) Simultaneous estimation of pose and motion at highly dynamic turn maneuvers. In: *Proceedings of the 31st DAGM symposium*, Springer, Heidelberg, pp 262–271
20. Schmidt J, Wöhler C, Krüger L, Gövert T, Hermes C (2007) 3D scene segmentation and object tracking in multicocular image sequences. In: *Proceedings of the fifth international conference on computer vision systems*, Bielefeld. <http://biecoll.uni-bielefeld.de/volltexte/2007/29/>
21. Fielding G, Kam M (1997) Applying the Hungarian method to stereo matching. In: *Proceedings of the IEEE conference on decision and control*, San Diego, pp 549–558
22. Stein F (2004) Efficient computation of optical flow using the census transform. In: *Proceedings of DAGM annual pattern recognition symposium*, Tuebingen, pp 79–86
23. Franke U, Joos A (2000) Real-time stereo vision for urban traffic scene understanding. In: *Proceedings of the IEEE intelligent vehicles symposium*, Dearborn, pp 273–278
24. Bock HH (1974) Automatische Klassifikation. Vandenhoeck & Ruprecht, Göttingen
25. Barrois B, Wöhler C (2008) Spatio-temporal 3D pose estimation of objects in stereo images. In: *Proceedings of the sixth international conference on computer vision systems*, vol 5008, Lecture notes in computer science. Springer, Heidelberg, pp 507–516
26. Duda RO, Hart PE, Stork DG (2000) Pattern classification, 2nd edn. Wiley, New York
27. Besl PJ, McKay ND (1992) A method for registration of 3-d shapes. *IEEE Trans Pattern Anal Mach Intell* 14(2):239–256
28. Zhang Z (1992) Iterative point matching for registration of free-form curves. IN-RIA, Technical Report. <http://ftp.inria.fr/INRIA/publication/publi-ps-gz/RR/RR-1658.ps.gz>
29. Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) Numerical recipes in C: the art of scientific computing. Cambridge University Press, New Delhi
30. Horn BKP (1987) Closed-form solution of absolute orientation using unit quaternions. *J Opt Soc Am A* 4(4):629–642
31. Hamilton WR (1866) Elements of quaternions. Longmans, Green, London
32. Hartley R, Zisserman A (2003) Multiple view geometry in computer vision, 2nd edn. Cambridge University Press, Cambridge
33. Krüger L, Wöhler C (2011) Accurate chequerboard corner localisation for camera calibration and scene reconstruction. *Pattern Recognit Lett* (to appear)
34. Hermes C, Einhaus J, Hahn M, Wöhler C, Kummert F (2010) vehicle tracking and motion prediction in complex urban scenarios. In: *Proceedings of the IEEE intelligent vehicles symposium*, San Diego
35. Käfer E, Hermes C, Wöhler C, Ritter H, Kummert F (2010) Recognition of situation classes at road intersections. In: *Proceedings of the IEEE international conference on robotics and automation*, Anchorage
36. Barrois B, Hristova S, Wöhler C, Kummert F, Hermes C (2009) 3D pose estimation of vehicles using a stereo camera. In: *Proceedings of the IEEE intelligent vehicles symposium*, Xi'an
37. Wöhler C (2009) 3D computer vision. Springer-Verlag, Berlin/Heidelberg

Tidal Energy

IAN G. BRYDEN

Head of Institute for Energy Systems, School of Engineering, The University of Edinburgh, Edinburgh, UK

Article Outline

Glossary

Definition of the Subject

Introduction

The Tides

Harnessing the Energy in the Tides

Future Directions

Bibliography

Glossary

Barrage An artificial barrier, usually constructed across an estuary, designed to maintain different water levels on either side.

Ebb tide The stage of a tide in which the water depth is reducing.

Flood tide The stage of a tide in which the water depth is increasing.

Neap tide Smaller fortnightly tides falling between the spring tides.

Spring tide Large fortnightly tides, resulting from the Earth, Sun, and Moon being approximately in line.

Tidal basin The area of water behind a tidal barrage.

Tidal currents Water currents resulting from tidal action.

Tidal stream A tidal current which has been accelerated as a result of the natural geography.

Tides The response of the world's seas and oceans to the influences of the Moon and the Sun relative to the Earth.

Turbine A mechanical system in which energy in a moving fluid is transferred to a rotating system.

Definition of the Subject

Tidal energy, as interpreted in this essay, is considered to be the artificial extraction of energy from: either the

rise or fall of the sea surface under the influence of tides or the extraction of energy from tidally driven currents. The associated theoretical energy resources are considerable on a global scale, but the geographic conditions necessary for significant tidal ranges, or current velocities, do tend to be restricted to a relatively small number of sites worldwide. Some of the most attractive tidal range sites, however, such as the Severn Estuary, between England and Wales, and the Bay of Fundy in Canada possess very considerable energy flux densities. Similarly, the most energetic tidal currents, such as in the Pentland Firth, to the north of the Scottish mainland, would also appear to offer major prospects for development.

Although there has been considerable progress in recent years toward the commercial exploitation of the tides for energy, the use of tidal energy is not new. Tide “mills” were used in northern France and southern England as far back as the late Roman era but were most prolific during medieval times. Tidal waters turned water wheels to drive cereal grinding apparatus in areas with few rapidly flowing rivers. The Eling Tide Mill, for example, is still operational, largely as an educational and tourist facility, and is a very early example of tidal entrainment, represented by a barrage positioned to interfere with natural tidal flows. Entrainment is a more general term than “barrage” and encompasses alternative techniques such as lagoons. In the twentieth and twenty-first centuries, tides have been seriously reexamined as sources of energy to power industry and commerce.

Introduction

Despite the installation of a major tidal energy plant at La Rance, France, in the 1960s, research and development throughout the 1970s and 1980s was largely restricted to university laboratories, with little commercial involvement. Interest in the use of the tides for the generation of electricity has increased significantly, however, in the last 20 years. From the early 1990s to the present day, governments have started, once again, to consider the tides as a potential source of energy and industry has seen prospects for profitable development. Governmental interest has largely been driven by concerns over long-term energy security and climate change. The deployment of strategic tidal capacity remains, however, a significant challenge,

which will require significant effort and resolve if the expectations of politicians and industrialist are to be achieved by 2020 and beyond.

The Tides

Explaining the existence of tides, using Newton's theories of gravitation and the mechanics of motion, represented a major step toward understanding the universe around us. The "equilibrium tidal theory," which is sometimes known as "the Newtonian theory of tides," gives a partial description of tidal behavior for an abstract water covered planet Earth and is outlined in most introductory texts on oceanography [1]. In effect, the tides represent a marine manifestation of energy fluxes present in the Earth–Moon–Sun system. These fluxes are complicated by the presence of land masses, which can result in local enhancement of the energy fluxes, where conditions are appropriate. The Bristol Channel between England and Wales and the Bay of Fundy in Canada are two noteworthy examples, which possess exceptionally high tidal ranges.

Although tides and tidal influences can be predicted centuries in advance because of the essentially mechanistic effects which drive them, the sea surface is also subject to weather-related effects of which waves are the most obvious. In addition, pressure effects and wind-driven currents will be superimposed upon tidal effects. These can, however, now be forecast over periods of hours and days from meteorological information.

Harnessing the Energy in the Tides

Approaches to Exploitation

There are two approaches to harnessing energy from tides:

- The first is to exploit the rise and fall of the sea level through entrainment. This includes "traditional" barrage methods as well and tidal lagoons and fences.
- The second approach involves harnessing local tidal currents using technology reminiscent of wind turbines.

Tidal Barrages

Principles and History Sites of particular interest include the Bay of Fundy in Canada, which has

a mean tidal range of 10 m, the Severn Estuary between England and Wales, with a mean tidal range of 8 m, and the Rance Estuary in Northern France, with a mean range of 7 m. A tidal barrage power plant has been operating at La Rance in Brittany since 1966 [2]. This plant is capable of generating 240 MW and incorporates a road crossing of the estuary. It has recently undergone a major 10-year refurbishment program.

Barrages have also been constructed at Annapolis Royal in Nova Scotia (18 MW), the Bay of Kislaya, near Murmansk (400 kW), and at Jangxia Creek in the East China Sea (500 kW) [3]. Schemes have been proposed for the Bay of Fundy and for the Severn Estuary but have never been built.

The prospect of generating electricity in the Severn Estuary has excited engineers and planners since the end of the nineteenth century, even before there was a significant demand for electricity! A serious proposal for an 800-MW generation scheme was made in 1925 [4], and this particular option continued to be examined seriously into the 1950s. Although the UK government has recently been considering options for the commercial development of the Severn Estuary, this has now been rejected for the foreseeable future.

In South Korea, the Sihwa Lake Tidal Power Plant was completed in 2010 with a preliminary capacity of 254 MW and has replaced La Rance as the largest tidal entrainment system. The plant uses ten 25.4 MW submerged bulb turbines. Unlike La Rance, the plant operates in flood rather than ebb mode in order to minimize some very specific environmental effects in the site.

Tidal barrages can operate in a variety of modes. These can be broken down initially into single basin schemes and multiple basin schemes. The simplest of these are the single basin schemes:

Single-Basin Tidal Barrage Schemes These require a single barrage across an estuary, as shown in Fig. 1. There are three different methods of generating electricity with a single basin. All involve a combination of sluices which, when open, allow water to flow relatively freely through the barrage and/or gated turbines, which generate electricity when required.

Ebb Generation: Electricity is generated only during the ebb tide. During the flood tide, water is allowed to flow freely through sluices into the barrage.

At high tide, the sluices are closed and water retained behind the barrage. When the water outside the barrage has fallen sufficiently to establish a significant head between the basin and the open water, the basin water is allowed to flow out through low head turbines and to generate electricity. The system behavior can be considered in phases. These are represented in Fig. 2 to show the periods of generation within the tidal cycle.

Typically the water will only be allowed to flow through the turbines once the head is approximately half the tidal range. This method will generate electricity for, at most, 40% of the tidal range.

Flood Generation: The sluices and turbine gates are initially kept closed during the flood tide to allow the water level to build up outside of the barrage, as shown in Fig. 3. As with ebb generation, once a sufficient head has been established, the turbine gates will

be opened and water can flow into the basin generating electricity.

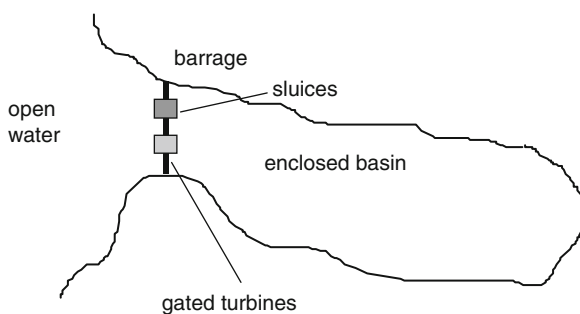
This approach is generally viewed less favorably than the ebb method, as keeping a tidal basin at low tide for extended periods might have detrimental effects on the environment and shipping. However, the special circumstances in the Korean Sihwa project have made flood mode the preferred option.

Two-Way Generation: It is possible, in principle, to generate electricity in both ebb and flood. Unfortunately, computer models do not generally suggest a major increase in the energy production. In addition, there would be further costs associated with requirements for either two-way turbines or a double set to handle the two-way flow. Advantages include, however, reduced periods with no generation and peak power would be lower, allowing reduction in generator costs.

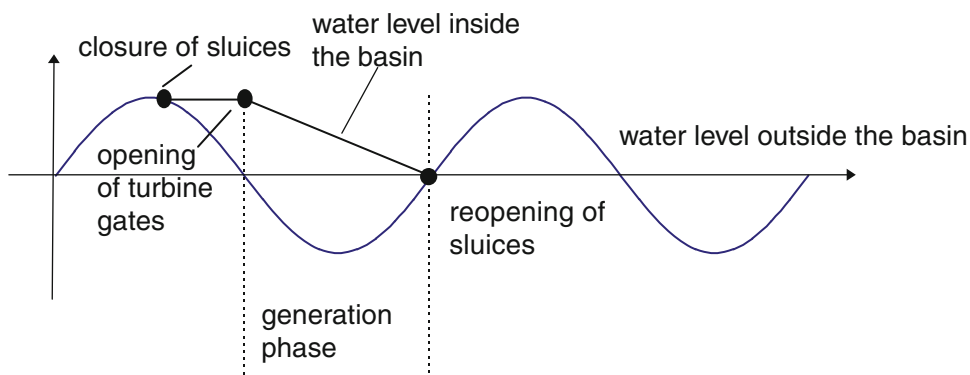
Double Basin Systems Single basin systems can only ever deliver energy during the parts of the tidal cycle. Double basin systems, as shown schematically in Fig. 4, might allow energy storage and enhanced control over output power levels.

The main basin might behave like conventional ebb or flood generation single basin system. Some of the electricity generated could be used to pump water to and from the second basin to ensure a more robust generation capability.

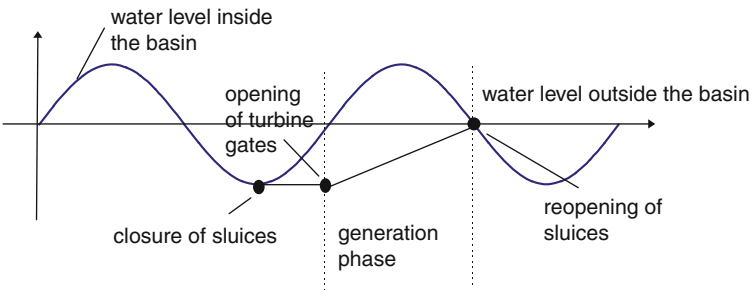
It is not likely that multiple basin systems will become popular, as low head turbines are unlikely to be sufficiently efficient to allow economic storage of energy.



Tidal Energy. Figure 1
Hypothetical tidal barrage configuration



Tidal Energy. Figure 2
Water levels in an ebb generation scheme



Tidal Energy. Figure 3
Water levels in a flood generation scheme

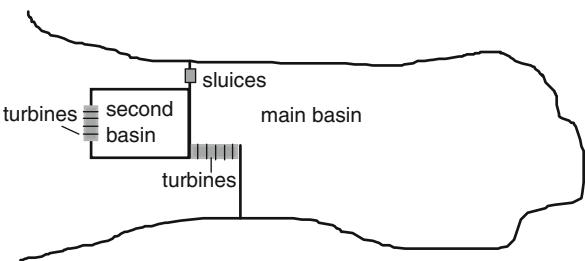
The overall efficiency of such low head storage is unlikely to exceed 30%. Conventional pump storage systems, with overall efficiencies exceeding 70%, are more attractive, especially as they are proven technology.

Possible Sites for Future Tidal Barrage Developments
Worldwide there are numerous sites which might be technically suitable for development, although whether the resource can be developed economically is yet to be conclusively determined [3]. These include, and this is not a definitive list:

Site	Mean tidal range (m)	Barrage length (m)	Estimated annual energy production (GWh)
Severn Estuary (UK)	7.0	17,000	12,900
Solway Firth (UK)	5.5	30,000	10,050
Bay of Fundy (Canada)	11.7	8,000	11,700
Gulf of Cambay (India)	6.1	25,000	16,400

Tidal Lagoons

Tidal barrage developments are likely to cause significant environmental change. Whether this would be acceptable would depend upon the individual development



Tidal Energy. Figure 4
Hypothetical two basin system

circumstances. Ebb generation will, for example, result in raised mean water levels. A barrage will also cause obstruction to shipping and other maritime activities, which will have an effect on the local economy.

Artificial lagoons [5] have been proposed as alternatives to barrages. The principal advantage is that the coastline, including the intertidal zone, would be largely unaffected. Careful positioning of the lagoon could also ensure that shipping routes would be unaffected. Although preliminary studies suggest that costs might be competitive with other sources of renewable energy, there has not yet been any in-depth peer-reviewed comparative assessment of the tidal lagoon concept.

Tidal Current Technology

Principles and History The development of a barrage or lagoon would represent a substantial investment of time and money. Many planners and engineers favor the development of tidal current systems, which could be developed incrementally. Indeed,

the progressive development of a local tidal current resource might allow subsequent advances in technology, or understanding of the resource, to be incorporated at later stages. The most significant early attempt to prove the practicality of tidal current power was conducted in Loch Linnhe in the Scottish West Highlands in the early 1990s [6]. This used a turbine held mid-water by cables, which stretched from a seabed anchor to a floating barge.

In 2000, a large floating device (the ENERMAR project [7]) was tested in the Strait of Messina between Sicily and the Italian mainland. Between May 2003 and October 2009, Marine Current Turbines (MCT) Ltd [8], of Bristol, England, demonstrated a 300-KW pillar-mounted prototype system called SeaFlow in the Bristol Channel.

Figure 5 shows the SeaFlow system with its nacelle raised into the “maintenance position.” This system, having completed its test program, has now been decommissioned.

The Canadian Clean Current project involved the testing of a ducted horizontal axis machine at Race Rock between July and September 2006, before being

removed in May 2007 and subsequently reinstalled after modification in October 2008.

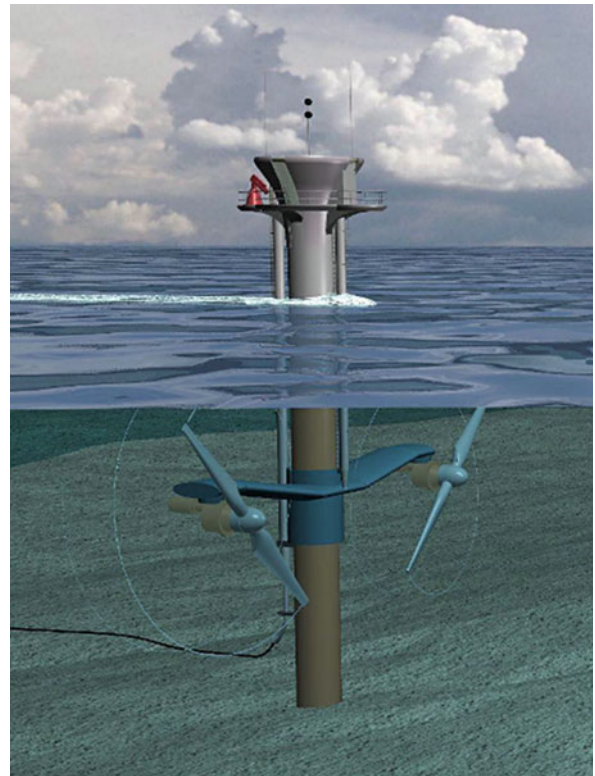
MCT installed their commercial scale prototype, SeaGen, in Strangford Narrows in Northern Ireland in 2008. Although having some similarities with SeaFlow, it is equipped with two rotors and has a rated capacity of 1.2 MW. An artist’s impression is shown in Fig. 6, and the operational system is shown in Fig. 7.

In Norway, the Hammerfest Strom system [9] demonstrated between 2003 and 2007 that a 300-KW pillar-mounted horizontal axis systems could operate in a fjord environment. The company now has ongoing plans for a larger 1-MW system, which it is intended to be available for commercial installation from 2012.

In the USA, Verdant Power Ltd. successfully demonstrated an array of six 5-m-diameter horizontal axis tidal turbines in New York’s East River from 2006–2008 [10]. The company now has detailed plans for a substantially larger development, using larger, more advanced technology.



Tidal Energy. Figure 5
SeaFlow with the nacelle raised



Tidal Energy. Figure 6
Artist's impression of SeaGen



Tidal Energy. Figure 7
Photograph of SeaGen in operation

In 2007, The European Marine Energy Centre [11], which was established in 2004 to allow the testing of full-scale marine energy technology in a robust and transparent manner, became fully equipped for the testing of tidal, as well as wave energy technology. The tidal test berths are located off the south-western tip of the island of Eday, in an area known as the Fall of Warness. The facility offers five tidal berths at depths ranging from 25 to 50 m in an area 2 km across and approximately 3.5 km in length. Each berth has a dedicated cable connecting back to the local grid. At the time of writing, there is one fully operational device [12] installed.

This is operated by Open Hydro Ltd. and is a novel annular turbine system held by twin vertical pillars. The system can be seen in its maintenance position in Fig. 8. The company has also been testing a novel barge-based installation system which could significantly lower costs.

There are presently two additional devices in the process of being installed at EMEC. The first, shown in Fig. 9, has been developed by Tidal Generation Ltd.

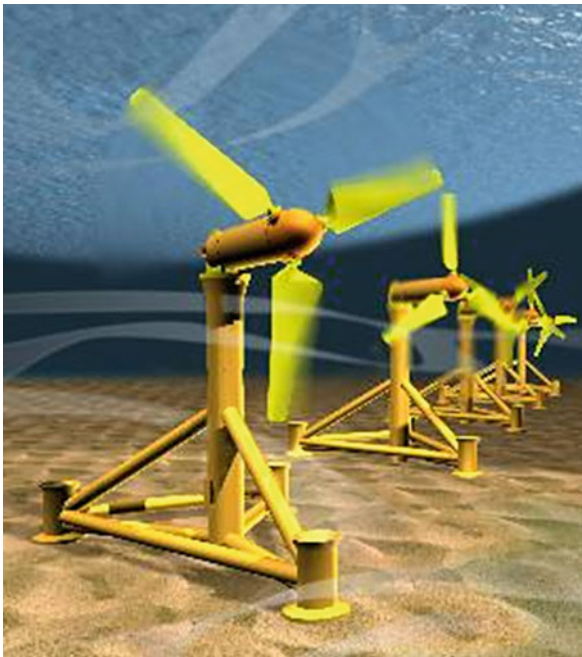
(TGL) and the second by Atlantis Power Ltd. Although EMEC was the world's first dedicated marine energy testing center, there are now ambitious developments elsewhere in the world, notably in Canada, where FORCE (Fundy Ocean Research Centre for Energy) is developing a significant test capability.

The physics of the conversion of energy from tidal currents is superficially very similar to the conversion of kinetic energy in the wind. Many devices have, therefore, a resemblance to wind turbines. There is, however, no total agreement on the form and geometry of the technology itself. Wind systems are generally horizontal axis rotating turbines with the axis of rotation parallel to the direction of flow. Many tidal developers also favor this approach, as can be seen in the images of prototype systems. Vertical axis systems, in which the axis of rotation is perpendicular to the direction of flow, have not, however, been rejected. ENEMAR, for example, used a vertical axis turbine.

Energy Available in Tidal Currents It is usual in renewable energy to consider the theoretical, technical, and practical resource. The theoretical resource is that



Tidal Energy. Figure 8
The Open Hydro System installed at EMEC



Tidal Energy. Figure 9
Artist's impression of the TGL Ltd. system being installed at EMEC

which could in principle be extracted, without consideration of technology or constraints. The “technical” resource is what could be exploited using reasonably available technology. The “practical” resource is what could be exploited after consideration of external constraints, competing use (shipping lanes, etc.) and environmental sensitivity.

It is tempting to draw analogies between wind power and tidal current power when assessing the resource. In both, it is possible to calculate the kinetic energy carried by a moving fluid and to relate this to a kinetic energy flux density. This, however, represents only part of the energy flux in a current: Energy is dissipated through “friction” between the water and boundaries; the surface of the sea in current will not be in level, and there will be changes in the potential energy of the water as it flows across a site. Generally, the potential energy changes and the frictional dispersion will exceed the kinetic flux!

Any serious analysis of the development potential of a tidal site should also consider the temporal and spatial complexity of the flow and the impact of extraction on the flow patterns. This will generally involve

numerical modeling of both the tides and the technology to allow assessment of the impact for proposed extraction scenarios, which will require high quality tidal/hydraulic data.

The “theoretical” resource will be difficult to identify for all but the simplest environments. It will usually be easier to estimate the “practical” resource by identifying what would be acceptable changes in the flow environment and using numerical methods to determine how much energy could be extracted without exceeding these limits. Individual prospective tidal sites, either estuaries or high-flow-speed areas, need site-dependent consideration. This makes the production of high-level resource assessments very difficult. Knowledge of the processes is now sufficient, however, to allow robust estimates of the practical resource at a selected site, provided sufficiently detailed data is available.

Development Options for Tidal Currents Tidal devices operate in a very different environment from onshore wind turbines, with specific problems associated with installation, survivability, and maintenance. These will need to be solved before true commercial exploitation can be achieved. Development options often involve the use of dedicated installation and maintenance vessels, which suggests that tidal currents might only be economically exploited in large sites, justifying the use of expensive infrastructures.

Many industrial, commercial, and public bodies have suggested that there is a high degree of synergy between a prospective tidal current generation industry and the offshore oil and gas industry. This offers the prospect of a new industry developing in partnership with the petroleum sector, which could, perhaps, result in accelerated development.

Tides are essentially predictable as they derive from astronomic processes, as discussed earlier in this article. Wind power systems are dependent upon random atmospheric processes and can be difficult to integrate into strategic electricity distribution networks. The predictability of the tides might make such integration much easier.

Prototype devices are now available but there are still issues to be resolved before commercial exploitation, for example: there is only limited understanding of how devices will operate in arrays, although working theories are becoming available; turbulence may

also be a problem as extreme fluctuations often exceeding 10% of the time averaged mean have been measured [13]. System designs must accommodate this; similarly, tidal devices will be exposed to waves, and this will also pose design challenges, if the systems are to be robust.

Such gaps in understanding should not, however, prevent ongoing deployment of technology, provided that developers are aware of the constraints that knowledge gaps impose and recognize that they themselves are part of the research process, which will ultimately allow cost-effective exploitation of the tidal current resource.

Future Directions

In a future in which energy costs are likely to rise, assuming that low-cost nuclear fusion or other long-term alternatives do not make unexpectedly early arrivals, tidal barrage schemes could prove to be major providers of strategic energy in the mid-twenty-first century and beyond, although the recent decision by the UK Government not to proceed with the development of a Severn barrage has disappointed many. The technology for tidal barrage systems is already available, and there is no doubt, given the experience at La Rance, that the resource is substantial and available. Ongoing development of the Korean resource also suggests that future development elsewhere in the world is possible. The Bay of Fundy remains a tantalizing option for a future in which energy costs are likely to rise considerably in real terms.

It is likely that tidal current systems will continue to be developed. The first truly commercial developments should appear this decade if progress continues. Tidal current systems may not have the strategic potential of barrage systems, but, in the short term at least, they do offer opportunities for supplying energy in rural coastal and island communities. In the longer term, massive sites such as the Pentland Firth could become strategically important. In November 2008, The Crown Estate, which owns the seabed in the Firth, invited proposals for the development of sites in the firth and surrounding waters [14]. Leases have now been granted for 1.6 GW of installed capacity, covering both wave and tidal current development. Of this, some 1.2 GW is intended to be tidal.

Bibliography

1. Bearman G (ed) (1997) Waves, tides and shallow water processes. The Open University, Milton Keynes
2. Banal M, Bichon A (1981) Tidal energy in France. The Rance tidal power station-some results after 15 years in operation. In: Proceedings of the second international symposium on wave and tidal energy, Cambridge
3. Boyle G (ed) (1996) Renewable energy. The Open University, Milton Keynes
4. "Sewern Barrage Scheme", Hansard, 4th February 1926
5. www.tidalelectric.com
6. www.itpower.co.uk/ResearchDevelopment.htm
7. www.pontediarchimede.com
8. www.marineturbines.com
9. www.tidevannsenenergi.com
10. www.verdantpower.com
11. www.emec.org.uk
12. www.openhydro.com
13. Norris J, Bryden I (2007) The European marine energy centre (EMEC): facilities and resources. Proc Inst Civil Eng 160(2): 51–58. doi:10.1680/ener.2007.160.2.51
14. www.thecrownstate.co.uk/newscontent/92-round-1-pentland-firth.htm

Topographic Solar Radiation Modeling for Environmental Applications

HOFIERKA JAROSLAV

Department of Geography and Regional Development,
University of Presov, Presov, Slovakia

Article Outline

Glossary

Definition of the Subject

Introduction

Modeling Topographic Effects on Solar Radiation

Computing Clear-Sky Radiation

Computing Real-Sky Radiation

The r.sun Solar Radiation Model

Computing Spatially Distributed Solar Radiation
Databases

Future Directions

Acknowledgments

Bibliography

Glossary

Geographic information system (GIS) Any information system that is used to collect, store, process, and present geographic information. In a narrow sense it means a specific application software dealing with geographic information. In a broader sense, it includes software, hardware, data, and people.

Digital elevation model (DEM) A digital representation of the land surface expressing a spatial distribution of elevations. It can be represented by a raster or vector data model.

Topography A study of Earth's surface geometry. It is also known as geomorphometry, a science of quantitative land surface analysis. The land surface (relief, terrain) is characterized by a set of morphometric parameters including slope, aspect, curvatures, and other more complex parameters.

Solar radiation database A set of digital maps representing a spatial and temporal distribution of solar radiation and related meteorological data over a specific region.

Clear-sky radiation Solar radiation under clear-sky atmospheric conditions. The extraterrestrial solar radiation is attenuated only by air mass with atmospheric gases and aerosols.

Real-sky radiation Solar radiation under real-sky conditions, that is, including the cloudiness.

Linke turbidity factor A measure of solar beam radiation attenuation caused by water vapor and aerosol particles in the atmosphere.

Raster A GIS data model representing geographic phenomena as arrays of cells that store attribute values.

Viewshed A graph showing the angular distribution of sky visibility and obstruction at the location. It is calculated from a DEM by searching in a specified set of directions around the location.

Definition of the Subject

Solar radiation exhibits strong spatial and temporal variations influenced by many factors. For example, basic patterns of seasonal and daily variations given by astronomic factors are modified by actual atmospheric conditions influencing absorption, scattering, and solar radiation spectrum. Such data offer some basic information about the available solar radiation. However,

patterns at regional and local scales may significantly deviate from these general trends. Large land surface features may cause strong local radiation gradients, thus inducing specific microclimate and hydrological conditions. Consequently, in mountainous regions, topography plays a critical role in the spatial distribution of solar radiation incident on the surface.

Currently available solar radiation databases derived from ground or satellite measurements contain data with an insufficient level of detail that in most cases do not represent adequately the spatial variability in solar radiation caused by topographic factors. To account for this variability, topographic solar radiation models, often integrated with geographic information systems (GIS), have been developed. Topographic factors are derived from digital elevation models (DEMs) representing the land surface. Higher spatial resolutions in DEMs bring the necessary level of detail in local estimates of solar radiation and its components incident on the surface. Moreover, topographic models are capable of providing estimates of solar radiation and related solar parameters at various time horizons and periods necessary for many environmental applications.

Introduction

Solar radiation incident at any location on the Earth's surface is the result of complex interactions of energy between the atmosphere and land or ocean surfaces. On a global scale, the latitudinal gradients of radiation are caused by the Earth's geometry and its rotation and revolution around the Sun. At regional and local scales over land, surface heterogeneities constitute the major factor modifying the spatial distribution of radiation. Spatial variabilities in land surface elevation, slope (inclination), aspect (orientation), and shadows cast by terrain features create strong local gradients that modify the global radiation field.

Spatial and temporal variations in incoming solar radiation determine the patterns and dynamics of many environmental phenomena (e.g., air and soil temperature, water balance, snow melting, photosynthesis, evapotranspiration, or biomass production) with a direct impact on the natural distribution of certain plants. Therefore, this spatial variability should be taken into account when studying the surface energy

balance or when preparing the localization of production crops (from agriculture or forestry). Moreover, high-resolution spatially distributed solar radiation data are also desired for solar energy applications, including building design, landscape architecture, photovoltaics (PV), or remote sensing.

At present, a number of solar radiation databases are available, such as NASA's SSE (<http://eosweb.larc.nasa.gov/sse/>), the European Solar Radiation Atlas [1], MeteoNorm [2], Satel-Light [3], SOLEMI (<http://www.solemi.com/>), HelioClim-2 (<http://www.helioclim.net/>), RETScreen (<http://www.retscreen.net/>), PVGIS [4], or 3TIER (<http://www.3tier.com/>). These databases were developed from ground radiation measurements, satellite images with interpolation techniques, and solar radiation models. They provide various solar radiation data and other related information helpful in global and regional studies. However, the spatial resolution of these databases is usually not sufficient for local studies. For example, the typical spatial resolution of global/regional solar databases that are derived from the Meteosat second-generation meteorological satellites is 1–5 km. Local solar radiation studies with topographic effects require higher spatial resolutions (less than 1 km). The spatial distribution of solar radiation in mountainous regions is strongly influenced by topography, which cannot be captured by satellite-based estimates nor by sparse ground measurements using radiometers.

The importance of topographic distribution of solar radiation in hydrological and biological processes has long been known (e.g., [5–11]). In the northern hemisphere, south-facing slopes generally receive more solar radiation than northern slopes, and therefore have a different heat and water balance regime. Areas frequently shaded by terrain features (e.g., bottoms of deep valleys) exhibit different microclimate profiles than open space areas. To account for this variability, topographic solar models have been created and often integrated with geographical information systems (GIS). Such models provide rapid, cost-efficient, and accurate estimates of solar radiation over large regions, while considering slope inclination, aspect, and shadowing effects. Coupling radiation models with both GIS and image processing systems improves their ability to use various environmental data and to be combined with other environmental models.

Significant progress has been made toward developing solar radiation models in the last 2 decades [12]. One of the first GIS-based solar radiation models was SolarFlux [12, 13] developed for the Arc/Info GIS system commercialized by ESRI. A series of solar radiation algorithms were also implemented in the Genasys GIS using AML scripts [14]. To describe the underlining radiative processes, these models use rather simple empirical formulas and parameters that are spatially averaged (lumped) and therefore not suitable for assessments over large areas.

More advanced methods for ecological and biological applications are implemented in Solar Analyst, developed as an ArcView GIS extension [15]. In the preprocessing phase, the model generates an upward-looking hemispherical viewshed based on a digital elevation model (DEM). The solar model is suitable for local-scale studies. It is not flexible enough for calculations of atmospheric transmissivity or diffuse fraction of the total irradiance, since it only allows very limited settings for the available parameters, using either data from the nearest weather stations or just typical values. This makes its use too simplistic over large areas, with progressive deterioration of accuracy as the study area increases.

The SRAD model [16, 17] was designed to evaluate a complex set of short-wave and long-wave solar radiation interactions between the Earth's surface and atmosphere. Although based on a simplified representation of the underlying physics, the main solar radiation factors are considered and the model is able to characterize the spatial variability in landscape processes. However, it is specifically designed for the modeling of topo- and meso-scale processes, so that the calculation of solar radiation over large areas is also limited.

The *r.sun* solar radiation model has been developed by Hofierka [18] for the open-source environment of GRASS GIS [19]. The improved version of the model, later developed by Šúri and Hofierka [20], models all three components of solar radiation under both clear-sky and real-sky atmospheric conditions. The model has been successfully used in a wide range of applications including PV electricity production [4, 21].

Modeling Topographic Effects on Solar Radiation

Solar radiation is a sum of three components: beam radiation incoming directly from the Sun, diffuse

radiation from the sky (available even under overcast conditions), as well as radiation diffusively reflected from nearby terrain features. This reflected component exists only in non-flat areas. The local variation in solar radiation incident at the surface is determined by two major factors: topography and cloudiness. Topography affects the distribution of solar radiation fluxes via several topographic attributes: elevation, slope, aspect, and terrain features creating shadows. In the northern hemisphere, for example, south-facing inclined surfaces at high latitudes may receive a higher amount of solar radiation than flat areas.

Slope and aspect of the land surface modify the incident beam, diffuse, and ground-reflected components. These two parameters influence the spatial-temporal patterns of the radiation field over hilly terrain. The influence of slope and aspect is more pronounced under clear-sky (sunny) weather than under cloudy conditions, when diffuse radiation dominates. Shadows cast by neighboring terrain features may substantially reduce the amount of available radiation especially at low solar altitudes. Elevation affects the attenuation of radiation through the total optical thickness of the atmosphere. At high elevation, due to a lower optical thickness, the attenuation from atmospheric gases and aerosols is less pronounced, so that the incident radiation's magnitude is higher than at low elevation. Similar issues, albeit even more complex, exist when trying to simulate the radiation field over urban fabrics, due to the added complexity of rapid spatial variations in ground albedo, inter-reflections between buildings, complex shading issues, and additional attenuation due to pollution.

Most radiometric stations are equipped with sensors that measure global horizontal radiation, sometimes diffuse horizontal radiation, and more rarely direct normal radiation. If both the global and diffuse components are measured, the direct normal component can be calculated using the actual Sun altitude over the horizon, which can be calculated from time, date, and location coordinates. Similarly, the radiation incident on slopes can be estimated from global and diffuse radiation data, provided the ratio of global radiation on a slope to that on a horizontal surface is known [22]. Because there is usually a very sparse spatial coverage of meteorological stations measuring solar radiation, it is generally very useful to estimate the solar radiation components locally incident on the surface using a topographical

solar radiation model. Such a model may provide more accurate estimates than satellites especially in mountainous regions with strong topographic effects.

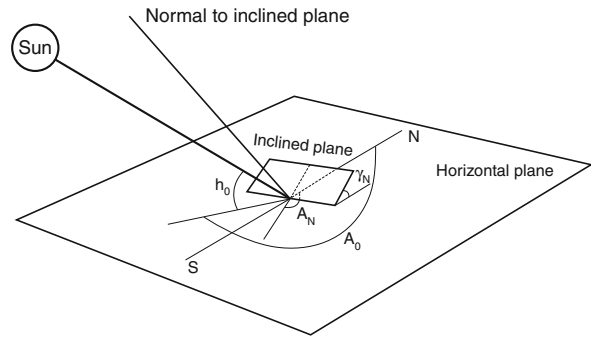
Digital Elevation Models

The most common digital representation of the land surface is via a DEM. In DEMs, elevations can be regularly spaced in the form of grids (a raster-based data model) or irregularly spaced as triangulated irregular networks (TINs). The cell of the grid (a minimal area with a unique value of elevation) defines the spatial resolution of the raster-based DEM. The DEMs based on the TIN data model are characterized by triangles with variable sizes.

DEMs are used to calculate slopes and aspects, and ultimately the amount of solar radiation incident at the surface. In mountainous regions, shadows cast by terrain features can greatly attenuate the available radiation, so that a shadowing analysis is necessary. However, the accuracy of such an analysis is very sensitive to the DEM's resolution. For applications at a global scale, solar radiation calculated with a spatial resolution of 10–40 km or more provides only overall information, which is not compatible with a detailed shadowing analysis. For resolutions of 1–5 km or better, the shadowing analysis should be considered, especially in mountainous or urban regions. In very local studies with DEMs at resolutions of hundreds to tenths of meters, the data accuracy strongly affects the shadowing analysis. In such cases, DEMs preserving the features of surface extremes (e.g., tops of the mountains, saddles, ridges, or bottoms of valleys) should be used rather than smoothed data. Moreover, digital surface models (DSM) that also include specific sun-obstructing features of the surface (e.g., forest, urban constructions, etc.) can be used. In the topographic modeling of solar radiation, DSMs are preferred because they better represent features casting shadows at lower solar altitudes. More information on DEMs, DSMs, data sources, production techniques, and quality issues can be found elsewhere, for example, [16, 23, 24].

Computing Clear-Sky Radiation

The Sun's position in the sky relative to the position of a point of interest at the surface is a key parameter in



Topographic Solar Radiation Modeling for Environmental Applications. Figure 1
Geometry of the solar beam

topographic solar radiation models. Basic solar geometry is defined by a set of solar parameters including latitude and longitude of the location, time, and solar declination. Then the position of the Sun in the sky with respect to a horizontal surface can be expressed by two coordinates: the solar altitude, h_0 , defined as the angle between the solar beam and the horizontal surface, and the solar azimuth, A_0 , defined as the horizontal angle between the Sun and the local meridian (Fig. 1).

The following description is for the solar topographic model that derived from the research by Šúri and Hofierka [20], which also resulted in the *r.sun* solar radiation model for the open-source environment of GRASS GIS [19]. Whereas the calculation of the beam component is quite straightforward, the main difference between all topographic models is in the treatment of the diffuse component [25]. This component depends on climate and regional surface conditions, and it is often the largest source of estimation error. The equations used in the *r.sun* model are based on the research conducted for the European Solar Radiation Atlas project [1, 26] and reflect European climate conditions. The ground-reflected radiation usually contributes only several percent to the overall global radiation on inclined surfaces and therefore is ignored by most models. However, larger contributions of this component may exist in mountains, especially in snow conditions. Nevertheless, the *r.sun* model takes also into account this component via simple equations that are adequate for most applications.

Beam Radiation

The beam irradiance on inclined surfaces, B_{ic} (W/m^2), is calculated as follows:

$$B_{ic} = B_{0c} \sin \delta_{exp} \quad (1)$$

or

$$B_{ic} = B_{hc} \sin \delta_{exp} / \sin h_0 \quad (2)$$

where B_{0c} is the beam irradiance normal to the solar beam (W/m^2), B_{hc} is the beam irradiance on horizontal surfaces (W/m^2), δ_{exp} is the solar incidence angle measured between the Sun and the inclined surface, and h_0 is the solar altitude.

If the inclined surface is defined by slope γ_N and aspect A_N , then the solar incidence angle δ_{exp} is defined as [27, 28]:

$$\sin \delta_{exp} = C'_{31} \cos (T - \lambda') + C'_{33} \quad (3)$$

where

$$C'_{31} = \cos \phi' \cos \delta, C'_{33} = \sin \phi' \sin \delta \quad (4)$$

and

$$\sin \phi' = -\cos \phi \sin \gamma_N \cos A_N + \sin \phi \cos \gamma_N \quad (5)$$

$$\tan \lambda' = -(\sin \gamma_N \sin A_N) / (\sin \phi \sin \gamma_N \cos A_N + \cos \phi \cos \gamma_N)$$

The hour angle T (rad) is calculated from the local solar time, t , expressed in decimal hours on the 24-h clock as:

$$T = 0.261799(t - 12) \quad (6)$$

The Sun's position given by the solar altitude h_0 and the solar azimuth A_0 is calculated as follows [27, 28]:

$$\sin h_0 = C_{31} \cos T + C_{33} \quad (7)$$

$$\cos A_0 = (C_{11} \cos T + C_{13}) / ((C_{22} \sin T)^2 + (C_{11} \cos T + C_{13})^2)^{1/2}$$

where

$$\begin{aligned} C_{11} &= \sin \phi \cos \delta, C_{13} = -\cos \phi \sin \delta, \\ C_{22} &= \cos \delta, C_{31} = \cos \phi \cos \delta, C_{33} = \sin \phi \sin \delta \end{aligned} \quad (8)$$

The hour angle of the time of sunrise/sunset over the horizontal plane, $T_h^{r,s}$, is defined as:

$$\cos T_h^{r,s} = -C_{33} / C_{31}. \quad (9)$$

The hour angle of the time of sunrise/sunset over the inclined surface, $T_i^{r,s}$, can be calculated by analogy:

$$\cos(T_i^{r,s} - \lambda') = -C'_{33} / C'_{31} \quad (10)$$

The value of the solar declination δ (rad) valid for a specified day can be found in astronomical handbooks or computed using one of the appropriate approximation formulas (see, e.g., [22, 29]).

The beam irradiance normal to the solar beam is attenuated by the cloudless atmosphere and calculated as:

$$B_{0c} = G_0 e^{-RmL} \quad (11)$$

where G_0 is the extraterrestrial irradiance normal to the solar beam (W/m^2) and the term RmL denotes the attenuation factor representing the Rayleigh optical thickness for a relative optical air mass m . The atmospheric turbidity is expressed by the Linke turbidity factor.

The beam irradiance on a horizontal surface B_{hc} (W/m^2) is given by the following expression:

$$B_{hc} = B_{0c} \sin h_0 \quad (12)$$

where h_0 is the solar altitude angle given by Eq. 7.

Diffuse Radiation

Calculations of diffuse radiation are always prone to some degree of approximation. Diffuse radiation depends on solar geometry, local topography, and atmosphere constituents causing absorption and scattering (e.g., clouds, particulates, aerosols). Diffuse radiation is anisotropic, that is, it varies over the sky depending on direction. Modeling anisotropy is a complex task, which would require knowledge of the amount of sky visible at a point (i.e., a sky-view factor) and of rapidly changing clouds at any instant. To simplify the task, it is often assumed that diffuse radiation is isotropic, which means that the diffuse sky radiance is constant over all directions. This can be a viable solution especially for long-term modeling.

Due to its complexity, the diffuse irradiance incident on the horizontal is usually obtained from

empirical models that have at least regional validity. The clear-sky diffuse radiation component of the *r.sun* solar radiation model presented here is based on the European Solar Radiation Atlas project [1].

The diffuse irradiance on a horizontal surface, D_{hc} (W/m^2) is estimated as the product of three terms: G_0 , a diffuse transmission function, T_n , dependent only on the Linke turbidity factor L , and a diffuse solar altitude function, F_d , dependent only on solar altitude h_0 [1]:

$$D_{hc} = G_0 T_n(L) F_d(h_0) \quad (13)$$

$F_d(h_0)$ reduces to 1 for a vertically overhead Sun. The product $G_0 T_n(L)$ then estimates the theoretical diffuse irradiance on the horizontal surface for the air-mass 2 Linke turbidity factor. The following second-order polynomial expression is used [1]:

$$T_n(L) = -0.015843 + 0.030543L + 0.0003797L^2. \quad (14)$$

The solar altitude function is evaluated using the expression:

$$F_d(h_0) = A_1 + A_2 \sinh_0 + A_3 \sin^2 h_0 \quad (15)$$

where the values of the coefficients A_1 , A_2 , and A_3 depend only on the Linke turbidity L .

The clear-sky diffuse irradiance on an inclined surface, D_{ic} (W/m^2) is estimated separately for sunlit, potentially sunlit, and shadowed surfaces. The exact expressions for these cases can be found in [30].

Ground-Reflected Radiation

The ground-reflected component of solar radiation is usually quite small (e.g., several percents of global radiation). It depends on topography and ground reflectance. The reflectance of the land surface is expressed via ground albedo. In flat areas, the ground reflected radiation is almost nonexistent; however, it may become significant in mountainous areas and higher latitudes with lower altitudes of the Sun and frequent presence of snow. Under such circumstances, the usual approach of considering the ground albedo as constant may lead to incorrect results. A good estimate of the albedo of all surrounding surfaces is a prerequisite to obtain correct results, particularly in calculations of the energy

balance of vegetation, amount of transpiration, and energy interception [22]. However, estimating the albedo value is not easy. Usually, tabulated (constant) values are used for typical land cover classes. In reality, the albedo may exhibit temporal variations due to ground cover and vegetation changes, variations in the soil's moisture content, periodic snow cover, etc. For example, Muneer [22] presents typical values of albedo for natural surfaces and ground or vegetation covers. In general, the values range from 0.15 to 0.30, but snow-covered ground could reach 0.9. Surface reflectance maps are generally derived from satellite spectral observations (e.g., [31–34]). Then these maps can be used to obtain the net solar radiation [32] over a large area. Spatial and temporal resolutions of the reflectance maps determine the accuracy of estimates. A higher spatial resolution of land cover maps may provide better spatial accuracy, whereas satellite spectral observation provides better estimates of temporal variations.

In the *r.sun* model, estimates of the clear-sky ground reflected radiation for inclined surfaces (R_i) rely on the isotropic assumption. The ground-reflected clear-sky diffuse irradiance received on inclined surfaces (W/m^2) in the absence of shading is proportional to the global horizontal irradiance, G_{hc} , mean ground albedo, ρ_g , and a fraction of the ground viewed by the inclined surface, $r_g(\gamma_N)$ [22]:

$$R_i = \rho_g G_{hc} r_g(\gamma_N) \quad (16)$$

where

$$r_g(\gamma_N) = (1 - \cos \gamma_N)/2 \quad (17)$$

and the global irradiance on the horizontal surface, G_{hc} (W/m^2) is given as a sum of its beam and diffuse component:

$$G_{hc} = B_{hc} + D_{hc}. \quad (18)$$

Computing Real-Sky Radiation

To properly estimate the real-sky irradiance, overcast and partially cloudy conditions must also be taken into account. However, human meteorological cloudiness observations are usually prone to subjective errors and do not sufficiently describe the physical nature or spatial and temporal patterns

of different types of cloud cover. The overcast irradiance (diffuse only) is calculated from clear-sky raster maps by application of a transmittance factor that parameterizes the attenuation of cloud cover. Examples of explicit calculations of this parameter can be found in [35, 36]. A simpler parameter, called the clear-sky index, can be used to calculate real-sky radiation for horizontal and inclined surfaces [20].

For the assessment of global radiation on a horizontal surface under overcast conditions, G_h , the clear-sky values, G_{hc} , are multiplied by the clear-sky index, k_c [3, 37, 38]:

$$G_h = G_{hc}k_c. \quad (19)$$

Index k_c represents the atmospheric transmission expressed as a ratio between the global horizontal irradiance under overcast and clear-sky conditions. For a set of ground meteorological stations, the clear-sky index can be calculated from the measured global irradiance, G_{hs} , and the computed values of clear-sky global radiation, G_{hc} :

$$k_c = G_{hs}/G_{hc}. \quad (20)$$

Alternatively, k_c can be derived from other sources of climatological data (e.g., cloudiness [39]). The raster maps of k_c must then be derived by spatial interpolation. The k_c index can also be calculated directly from modeled shortwave surface irradiance predictions derived from satellite data. The latter method is based on the similarity between the planetary albedo recorded by spaceborne radiometers and the surface radiant flux [3, 37, 40].

To evaluate the overcast global irradiance for inclined surfaces, G_i , the beam, diffuse, and reflected components are calculated similar to clear-sky equations by just substituting B_{hc} for B_h and D_{hc} for D_h . To compute D_h and B_h , the clear-sky index k_c has to be treated separately, as follows [20]:

$$D_h = D_{hc}k_c^d \quad (21)$$

$$B_h = B_{hc}k_c^b.$$

The diffuse fraction, or ratio of diffuse to global radiation D_h/G , is a strong function of cloudiness, and therefore is markedly different under clear and overcast skies. In Europe, the diffuse ratio typically varies

between 0.3 and 1.0 [40]. The underlying physical processes are quite complicated, and usually represented only by empirical equations [1, 39]. However, for many meteorological stations where the global irradiance, G_{hs} , is measured, the diffuse component, D_{hs} , is either measured or calculated from cloudiness, sunshine, or other climatological data. The raster map of D_{hs}/G_{hs} can be derived from such point values by spatial interpolation. Consecutively, the raster maps of diffuse and beam components of the clear-sky index can be computed as follows [20]:

$$D_h = G_h D_{hs}/G_{hs} \quad (22)$$

$$B_h = G_h - D_h$$

$$k_c^d = D_h/D_{hc} \quad (23)$$

$$k_c^b = B_h/B_{hc}$$

where subscript s is meant to distinguish data measured at meteorological stations (B_{hs} and D_{hs}) from the estimated values of B_h and D_h .

The *r.sun* Solar Radiation Model

The methodology described above has been applied to the *r.sun* solar radiation model within the open-source environment of GRASS GIS [18, 20, 41]. GRASS GIS is a free, open-source software that warrants users, developers, and researchers free access to the source code and ways to further improve it [19]. The *r.sun* module of GRASS GIS can be used for any geographical region and time to assess the solar radiation field, and evaluate the irradiance incident on any elemental surface. It has been also successfully applied to assessments of PV electricity production at regional and local scales [4, 20, 24]. A detailed description of the module can be found in [19] and in the GRASS GIS online manual (<http://grass.osgeo.org/>).

The *r.sun* model estimates the beam, diffuse, and reflected components of clear-sky radiation after taking into account the attenuation of extraterrestrial radiation by air-mass effects and atmosphere turbidity. The real-sky radiation is approximated from the latter results using the clear-sky index, which represents the attenuation of clear-sky radiation by cloudiness. The clear-sky index map is usually derived from ground radiation measurements by interpolation or from

satellite-based predictions. Calculations of the irradiance incident on inclined planes require the estimation of the beam and diffuse components of global horizontal radiation. Similar to the clear-sky index, the ratio of diffuse and global radiation can be derived by interpolating ground-based data or by models based on satellite data.

In summary, the *r.sun* module has the following key features [24]:

- It is a raster-based GIS program with spatially variable input and output data. Alternatively, selected parameters can be set as constants.
- It is a free, open-source program with available source code for further improvement.
- It provides all components of clear-sky and real-sky solar radiation for irradiation (Wh/m^2) and irradiance values (W/m^2).
- Its calculations can be performed assuming solar or civil time.
- It is scalable to various spatial resolutions and region sizes. Memory management and code optimization allows the use of high-resolution data.
- Integration with the open-source environment of GRASS GIS provides other GIS tools for processing the input and output data directly within a single computing environment.

The *r.sun* module works in two modes. *Mode 1* is for instantaneous calculations, for which raster-based maps of solar irradiance (W/m^2) and solar incident angles (degrees) are obtained. In *mode 2*, the raster-based maps provide daily sums of solar irradiation (Wh/m^2) and daily direct-sun duration (min). These are computed from the integration of irradiance values (derived in mode 1) that are calculated at a user-selected time step from sunrise to sunset. If such option is also selected, the computation can account for sky obstructions (shadowing) by local terrain features. These two modes can be used separately or in combination to provide estimates of incident radiation for any desired time step or interval. This is done by incorporating Unix shell scripts (e.g., in the Linux operating system).

Recently, the *r.sun* module has been supplemented by other GRASS GIS modules to help prepare spatially distributed solar databases [42]. For example, *r.sunyear* calculates the optimal tilt of south oriented planes in

order to maximize the collectable daily or annual irradiation with 1° precision. This module provides the optimized tilt of the collector plane, as well as the corresponding global irradiation values including the beam, diffuse, and reflected components.

The most computationally demanding part of topographic solar radiation modeling is the analysis of land surface shadows. To calculate shadows, the internal procedure of the *r.sun* module can be used. Alternatively, for large-size grids, the more efficient *r.horizon* module is preferable. The output of this module is a grid representing the elevation angles between the horizontal and the apparent horizon for any specified direction. Series of such grids describing the horizon angles over all directions (with a specified azimuthal step, usually between 3.75° and 15°) can then be used as the input to the *r.sun* module, thus simplifying its internal procedure. Both the *r.horizon* module and the internal shadowing algorithm of the *r.sun* module consider the curvature of the Earth and the angular distortions caused by various cartographic projections.

Computing Spatially Distributed Solar Radiation Databases

Solar radiation parameters are essential to many environmental processes and human activities. Thus it is important to determine with a sufficient spatial and temporal accuracy the amount of solar radiation intercepted by any point on the Earth's surface. Radiation data are usually available from selected ground meteorological stations operating under the auspices of the World Meteorological Organization (WMO). Such data can be retrieved from the World Radiation Data Centre (WRDC). The archives of the WRDC are available online (<http://wrdc-mgo.nrel.gov> and <http://wrdc-mgo.rssi.ru/>). However, such data do not cover areas between the measurement stations. In most cases, only global irradiation is available in these archives. Models must therefore be used to estimate the beam and diffuse components. This can be done by using additional meteorological data, such as cloud cover, the turbidity factor, and air temperature.

Currently, there are several satellites whose cloud data can be used to evaluate global radiation at regional or global scales. Various international and national

programs provide satellite-derived solar radiation maps, but their spatial resolution is usually coarse (5–300 km). Among such global initiatives, the International Satellite Cloud Climatology Project (ISCCP) and related programs such as NASA's Surface Solar Energy (SSE) are worth noting [43].

In several regions, digital solar atlases on CD or DVD have been developed to meet the increasing demand for accurate solar radiation data. These atlases are often supplemented by software to compute other parameters of interest and provide a graphical user interface to the maps. For example, the MeteoNorm software (<http://www.meteonorm.com>) contains various solar radiation or climatological data and software tools to perform spatial interpolations and estimate the irradiance on inclined surfaces. Another example is the European Solar Radiation Atlas [1] covering the European continent.

Currently, new web-based services provide new possibilities for solar radiation applications. For example, the SoDa Web service (<http://www.soda-is.com>) provides an integrated access to several online databases with a wide range of applications [43]. Other examples are the National Renewable Energy Laboratory (NREL) site (<http://rredc.nrel.gov/solar>), the RETScreen project (<http://www.retscreen.net>), the 3Tier web service (<http://www.3tier.com>), or the PVGIS site for PV assessments (<http://re.jrc.ec.europa.eu/pvgis/>).

These solar radiation databases provide basic solar radiation maps, such as mean daily values of global horizontal, direct normal, or diffuse horizontal radiation. Thus, the user with interests in the spatial distribution of solar radiation over land must use additional calculations and models to assess the effective solar irradiance/irradiation that is incident on tilted surfaces.

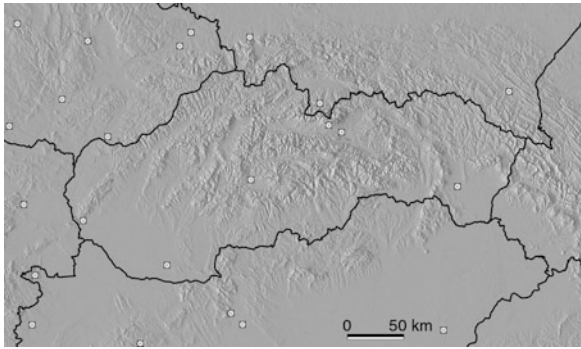
Recently, Hofierka and Cebecauer [24, 44] have developed the national solar radiation database for Slovakia with a spatial resolution of 100 m. This also includes topographic solar radiation maps suitable for environmental applications. The solar radiation database has been derived using the *r.sun* model described above, and additional solar radiation algorithms. The raster-based solar radiation database consists of 12 monthly mean and one annual mean of daily sums of global irradiation on a horizontal plane under real-sky and clear-sky conditions, spatially distributed values of optimum inclination angle of a plane maximizing the

solar radiation collection (e.g., for solar energy applications) and monthly means of real-sky global irradiation for optimally inclined planes. Moreover, several other raster-based maps were prepared, including the beam, diffuse, and reflected components of global irradiation, Linke turbidity coefficients, ratios of diffuse and global irradiation on a horizontal plane, ratios of global irradiation on optimally inclined and horizontal planes, as well as the underlying DEM.

Input Data

The data used in this study were derived from the existing data sources and then further processed in GRASS GIS and by the *r.sun* solar radiation model. The input dataset included a 100-m resolution DEM, measured values of global and diffuse radiation from 22 ground meteorological stations, and monthly maps of Linke turbidity coefficient.

As mentioned earlier, the essential input for topographic solar radiation modeling is a proper DEM. The spatial resolution of raster-based DEMs significantly affects the accuracy of modeling and spatial patterns of solar irradiation. The DEM used for the Slovakia database was based on the SRTM data from the C-band SAR interferometry instrument [45] acquired during an Endeavour space shuttle mission that took place in 2001. These data cover approximately 80% of the total Earth surface. In the northern hemisphere, the data covers all areas up to 60° latitude. The original resolution of the data is 1 arcsec (≈ 30 m). For areas outside the USA, only the coarser 3-arcsec data are freely available (<http://srtm.usgs.gov/>), which still determine an excellent resolution (≈ 90 m). The horizontal and vertical accuracy of the data is sufficient for most solar applications, even at local scales. In this study, the 3-arcsec data available in the WGS84 geographic coordinate system were reprojected and resampled to 100-m resolution (Fig. 2). To accurately take into account the shadowing effects of surrounding terrain, a 50-km strip of DEM behind the state border of Slovakia was also included. The detailed daily analysis of shadows on large grids throughout a year means very high computing demands, which can be minimized by the *r.horizon* module. A set of 96 data layers representing horizon angles with an azimuthal angle step of 3.75° were precomputed using this module.



Topographic Solar Radiation Modeling for Environmental Applications. Figure 2

Ground meteorological stations measuring solar radiation and DEM derived from the SRTM data

This considerably accelerated calculations in the *r.sun* module [44].

Ground meteorological stations measuring solar radiation and other related meteorological parameters (e.g., turbidity, cloudiness, etc.) are usually sparse and unevenly distributed. In most cases, their spacing is such that the spatial variability in solar radiation cannot be accurately derived. In Slovakia, only six meteorological stations measuring global solar radiation are available. To use the maximum possible information, these stations have been supplemented by another set of 16 nearby meteorological stations in neighboring countries, for a total of 22 stations used in this study (Fig. 2). The measurement of global and diffuse solar radiation is needed to assess the ratio of real-sky and clear-sky radiation represented by the clear-sky index and its beam and diffuse components. These are necessary in order to derive the real-sky radiation on inclined surfaces.

To prepare spatially distributed input data derived from measured or modeled solar radiation data, interpolation techniques are frequently used, such as splines, weighted averages, or kriging (see, e.g., [46–48]). However, a low spatial density of meteorological stations measuring solar radiation still requires a sophisticated interpolation approach, such as co-kriging [49, 50] or multivariate splines [20, 51], in order to include additional variables and useful information from a DEM or satellite images, and effectively improve the quality of interpolation. Using this approach, Šúri et al. [4] have prepared a PVGIS

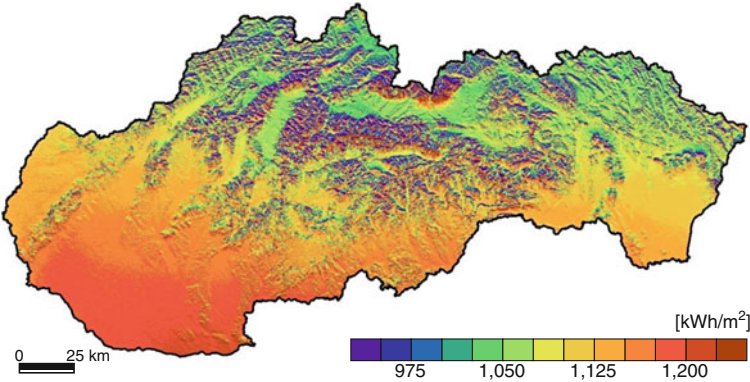
database (<http://re.jrc.ec.europa.eu/pvgis/>) with 1-km spatial resolution, which contains a comprehensive set of various solar radiation data. This includes the *r.sun* input parameters for Europe, which also covers the territory of Slovakia. The beam and diffuse clear-sky index components were derived from the monthly means of global and diffuse radiation measured at many meteorological stations (including the 22 stations considered in the Slovakia study) during the 1981–1990 period [1], and from clear-sky global irradiances computed using the *r.sun* model [4]. The input data for this study included the beam and diffuse clear-sky index components and Linke turbidity grid maps taken from the PVGIS database, which were all resampled to 100-m resolution using the *r.resamp.rst* interpolation resampling module in GRASS GIS. The original maps of monthly Linke turbidity factor were obtained from a global database [52] at their original spatial resolution of 5' (≈ 8 km). The estimation of turbidity is still prone to large uncertainties and affects the accuracy of clear-sky radiation maps, particularly those for the beam and diffuse components.

The monthly mean clear-sky index and diffuse fraction were interpolated using the *v.vol.rst* module in GRASS GIS, which is based on a tri-variate version of the Regularized Spline with Tension method [19]. Finally, the elevation variable was represented by the 100-m DEM derived from the SRTM data, thus enhancing the resolution and accuracy of interpolation in mountainous regions.

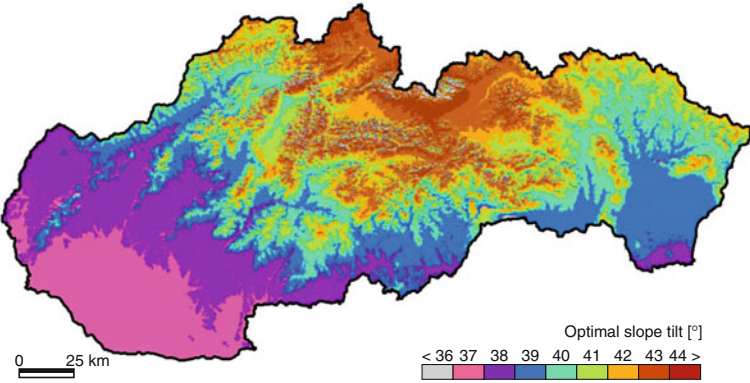
Calculation Procedure

The solar database represented by the final maps in Figs. 3–6 was computed by the *r.sun* module with the help of the above-mentioned GRASS GIS modules in three major steps [20]:

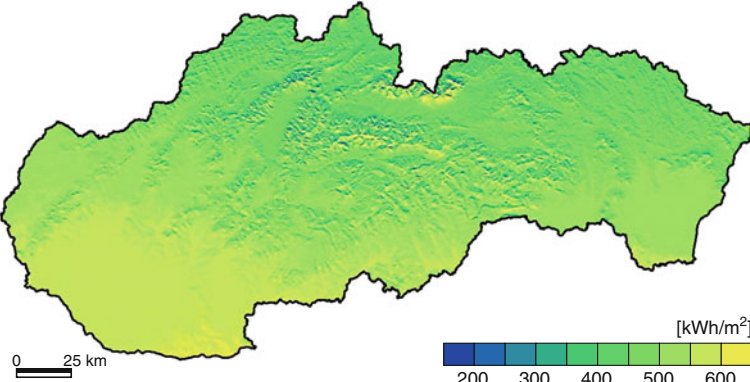
1. Computation of global clear-sky irradiation on a horizontal plane
2. Calculation of global real-sky irradiation on a horizontal plane and its beam and diffuse components using the beam and diffuse clear-sky index components
3. Computation of the optimal inclination of the south-oriented plane to maximize the annual irradiation yield using the *r.sunyear* module, and computation of irradiation for optimally inclined planes



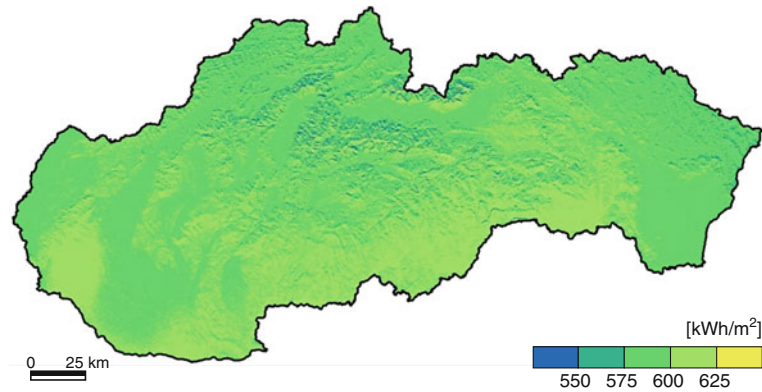
Topographic Solar Radiation Modeling for Environmental Applications. Figure 3
Annual real-sky global solar irradiation incident on the ground surface in Slovakia



Topographic Solar Radiation Modeling for Environmental Applications. Figure 4
Optimal slope tilt maximizing the annual solar input in Slovakia



Topographic Solar Radiation Modeling for Environmental Applications. Figure 5
Annual real-sky beam solar irradiation



Topographic Solar Radiation Modeling for Environmental Applications. Figure 6

Annual real-sky diffuse solar irradiation

All databases were derived in the form of spatially distributed raster-based maps representing monthly and annual means of daily sums of irradiation in kWh/m^2 . The long-term mean values were determined by the input data, that is, long-term monthly means of the clear-sky index and of the Linke turbidity coefficient. Although these important inputs have a coarse time resolution, the use of a high-resolution DEM provides the capability to better model the daily and seasonal dynamics of solar radiation. Values of slope and aspect were calculated from the SRTM DEM using the *r.slope.aspect* module. Calculations were performed for the median day of each month, taken as the effective average day. An integration time step of 15 min was used to obtain daily sums of irradiation. A shell-scripting technique in GRASS GIS was used to calculate monthly values of irradiation using the *r.sun* module in *mode 2*.

The resulting spatial maps of solar radiation consist of:

1. Monthly means of global irradiation including all components for clear and real skies on a horizontal plane
2. Monthly means of global irradiation including all components for real-sky incident at the land surface
3. Irradiance values incident at the land surface in 15-min steps for March, June, September, and December

The monthly means were subsequently summed up to get annual values.

The quality of outputs from any topographic solar radiation model reflects that of the source data and of the solar radiation methodology. A high-resolution DEM improves the accuracy of topography factors (absolute air mass and shadowing). The real-sky irradiation results are more influenced by the lack of representative ground-based measurements reflecting local variations and dynamics of cloudiness. In the Slovakia case, ground-based measurements were spatially interpolated using a sophisticated tri-variate interpolation method that can make the local variability of irradiation dependent on topography [20, 51].

The overall accuracy of the database represented by monthly and annual values of irradiation was assessed using basic statistics, the relative mean bias error (rMBE), and relative root mean squared error (rRMSE). The rMBE measures the overall bias. For the six available radiometric stations in Slovakia, it reached 3.1% and 2.7% for the monthly and annual real-sky global irradiations on a horizontal plane, respectively. The rRMSE, which is more a measure of random errors, was a bit higher – approaching 5.9% and 4.2% for the monthly and annual irradiations, respectively [24].

For Slovakia, the mean annual real-sky global irradiation incident on the surface (Fig. 3) averages from 925 kWh/m^2 on northern slopes and bottoms of valleys to 1,250 kWh/m^2 on southern slopes of mountains, with locally higher values. Interestingly, these southern slopes have a higher solar input than horizontal areas in the southern part of the country (1,150–1,200 kWh/m^2). The optimal slope tilt maximizing the annual solar irradiation

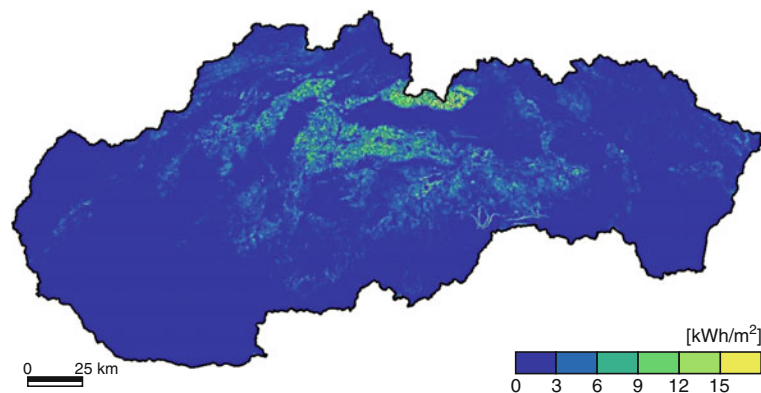
input is in the range of 36–44°, that is, slightly less than latitude (Fig. 4). This confirms the findings by Gueymard [53] for the US territory. The optimal slope tilt was calculated using the *r.sunyear* module that iteratively calculates the optimal slope tilt using the *r.sun* model [54]. The values of optimal slope tilt also explain why south-oriented, steeper slopes at higher latitudes may receive more radiation than horizontal areas at lower latitudes. Figure 4 also indicates that the spatial variability in global radiation incident on the ground is quite high in Slovakia, with mountainous areas exhibiting the strongest gradients, as could be expected.

Figure 5 shows the mean annual real-sky beam irradiation incident on the ground surface in Slovakia. This component is the most sensitive to changes in land surface geometry, especially under lower solar altitudes due to the occurrence of shadows. Mountains also often trigger a local increase in cloudiness, thus attenuating the beam radiation component. The mean annual beam irradiation varies from 100 kWh/m² on

northern slopes and valley bottoms to 650 kWh/m² on southern slopes and flat areas.

Diffuse radiation (Fig. 6) is the second most important component of global radiation. The amount of diffuse radiation is directly affected by the clear-sky index and to a lesser extent by topography. The *r.sun* model considers the terrain's slope, but not the indirect, sky blocking effect of the neighboring terrain on diffuse radiation. The mean annual diffuse irradiation varies only slightly from 500 kWh/m² on northern slopes and valley bottoms to 650 kWh/m² on southern slopes and higher elevations. Over flat areas, an intermediate value of ≈600 kWh/m² per year is typical. The spatial variability in diffuse irradiation is lower than in beam irradiation, since the influence of shading effects on this component is weaker.

Reflected radiation (Fig. 7) has the lowest share in global radiation, as could be expected. It exists only in non-flat areas, and reaches the highest values in mountains. The mean annual reflected irradiation is from



Topographic Solar Radiation Modeling for Environmental Applications. Figure 7

Annual real-sky reflected solar irradiation

Topographic Solar Radiation Modeling for Environmental Applications. Table 1 Mean, minimal, and maximal daily global irradiation incident on the ground surface in Slovakia (kWh/m²)

Daily irradiation (kWh/m ²)	Month												
	1	2	3	4	5	6	7	8	9	10	11	12	Year
Mean	0.9	1.6	2.7	3.9	4.8	5.0	5.3	4.4	3.2	2.2	1.1	0.7	3.0
Minimum	0.3	0.6	1.0	1.3	1.8	2.0	2.0	1.5	1.0	0.7	0.5	0.3	1.1
Maximum	2.3	3.2	4.1	4.7	5.5	5.8	6.1	5.3	4.4	3.5	2.3	1.7	3.9

0 kWh/m² in flat areas to 18 kWh/m² on steep slopes in mountains. The seasonal variability in albedo was not considered.

The seasonal variability in mean daily global irradiation over Slovakia is presented in Table 1. This variability is quite high due to the relatively high latitude (47°–49°). During the summer months (June to August), the mean daily irradiation is about 5.5 times higher than during the winter months (December to February). The highest seasonal variability and differences can be found on northern slopes (up to seven times), in contrast to southern slopes where differences are only ≈2.5 times. This seasonal variability is caused by topographic effects that are increased during winter when solar altitudes are lower and hence frequent shadows are cast by terrain features.

Future Directions

Topographic solar radiation models are increasingly available for various environmental applications. However, due to their complexity and input data requirements, they still lack some features that could improve the accuracy of the modeling results. While clear-sky beam component is well defined and easily calculated, the diffuse component is prone to approximations, even under clear-sky conditions. Thus the anisotropic character of diffuse radiation is often approximated by the isotropic assumption. The problem is even more complex under real-sky conditions. Spatial and temporal variability caused by clouds can overwhelm the topographic variability [12].

The long-term values of real-sky radiation incident at the land surface can be well approximated using ground-based or satellite-based measurements. The ground-based data are usually sparsely and unevenly distributed and therefore more advanced techniques of spatial interpolation might be necessary. The advanced interpolation and disaggregation techniques can be used to assess the real-sky radiation with higher spatial resolutions and topographic effects. However, the short-term changes away from weather stations lack adequate modeling tools. The satellite-based data are evenly distributed, however, often spatially averaged (smoothed) and therefore less accurate in mountainous areas. Still, topographic models and elevation data

can be used to disaggregate satellite data in areas with strong topographic effects, and thus increase the accuracy of estimates [55].

Recently, the *r.sun* topographic solar radiation model has been successfully applied to assessments of roof potential for solar energy applications in urban areas [21]. However, the analysis is limited to 2-D surfaces (roofs) and cannot be applied to vertical facades. Therefore, future improvements of topographic solar radiation models should also include extension to 3-D modeling for vertical surfaces including complex urban areas and overlapping plant canopies.

Acknowledgments

This work was supported by the Ministry of Education of the Slovak Republic under Grant VEGA 1/0355/09.

Bibliography

1. Scharmer K, Greif J (2000) The European Solar Radiation Atlas, vol 2, Database and Exploitation Software. Paris. Presses de l'École des Mines de Paris, France
2. Remund J, Kunz S, Lang R (1999) METEONORM: global meteorological database for solar energy and applied climatology, Solar Engineering Handbook, version 4.0. Meteotest, Bern
3. Hammer A, Heinemann D, Westerhellweg A, Ineichen P, Olseth JA, Skartveit A, Dumortier D, Fontoynt M, Wald L, Beyer HG, Reise Ch, Roche L, Page J (1998) Derivation of daylight and solar irradiance data from satellite observations. In: Proceedings of the 9th conference on satellite meteorology and oceanography, Paris, May 1998, pp 747–750
4. Šúri M, Huld TA, Dunlop ED (2005) PVGIS: a web-based solar radiation database for the calculation of PV potential in Europe. Int J Sustainable Energy 24:55–67
5. Cottle HJ (1932) Vegetation on north and south slopes of mountains in south-western Texas. Ecology 13:121–134
6. Wilson RG (1970) Topographic Influences on a Forest Microclimate. McGill University, Montreal, Canada
7. Halverson HG, Smith JL (1979) Solar radiation as a forest management tool: a primer of principles and application, General technical report PSW-33. The Forest Service of the US Department of Agriculture, Berkeley
8. Dozier J (1980) A clear-sky spectral solar radiation model for snow-covered mountainous terrain. Water Resour Res 16:709–718
9. Gates DM (1980) Biophysical ecology. Springer, New York, pp 96–146
10. Moore ID, Grayson RB, Ladson AR (1991) Digital terrain modeling: a review of hydrological, geomorphological and biological applications. Hydrol Process 11:47–54

11. Fu P, Rich PM (2000) A geometric solar radiation model with applications in agriculture and forestry. *Comput Electron Agric* 37:25–35
12. Dubayah R, Rich PM (1995) Topographic solar radiation models for GIS. *Int J Geogr Inf Syst* 9:405–413
13. Hetrick WA, Rich PM, Barnes FJ, Weiss SB (1993) GIS-based solar radiation flux models, American society for photogrammetry and remote sensing technical papers, vol 3, GIS photogrammetry and modeling. Wiley, New York, pp 132–143
14. Kumar L, Skidmore AK, Knowles E (1997) Modeling topographic variation in solar radiation in a GIS environment. *Int J Geogr Inf Sci* 11:475–497
15. Fu P, Rich PM (2000) The solar analyst 1.0 user Manual. Helios Environmental Modeling Institute, Lawrence, KS
16. Wilson JP, Gallant JC (2000) Secondary topographic attributes. In: Wilson JP, Gallant JC (eds) *Terrain analysis; principles and applications*. Wiley, New York, pp 87–132
17. McKenney DW, Mackey BG, Zavitz BL (1999) Calibration and sensitivity analysis of a spatially-distributed solar radiation model. *Int J Geogr Inf Sci* 13:49–65
18. Hofierka J (1997) Direct solar radiation modelling within an open GIS environment. In: *Proceedings of the joint European GI conference*, Vienna, pp 575–584
19. Neteler M, Mitasova H (2004) *Open Source GIS: a GRASS GIS approach*, 2nd edn. Kluwer, Boston
20. Šúri M, Hofierka J (2004) A new GIS-based solar radiation model and its application to photovoltaic assessments. *Trans GIS* 8:175–190
21. Hofierka J, Kaňuk J (2009) Assessment of photovoltaic potential in urban areas using open-source solar radiation tools. *Renewable Energy* 34:2206–2214
22. Muneer T (2004) *Solar radiation and daylight models*. Elsevier, Amsterdam
23. Hengl T, Reuter HI (2009) *Geomorphometry: concepts, software, applications*, vol 33, *Developments in Soil Science*. Elsevier, Amsterdam
24. Hofierka J, Cebecauer T (2008) Spatially distributed assessment of solar resources for energy applications in Slovakia. *Acta Facultatis Studiorum Humanitatis et Naturae Universitatis Prešovensis. Prírodné vedy Folia Geogr* 12:97–114
25. Perez R, Seals R, Ineichen P, Steward R, Menicucci D (1987) A new simplified version of the Perez diffuse irradiance model for tilted surfaces. *Sol Energy* 39:221–231
26. Rigollier Ch, Bauer O, Wald L (2000) On the clear sky model of the ESRA – European Solar radiation Atlas – with respect to the Heliosat method. *Sol Energy* 68:33–48
27. Krcho J (1990) Morfometrická analýza a digitálne modely georeliéfu. Bratislava, Veda (in Slovak)
28. Jenčo M (1992) Distribúcia priameho slnečného žiarenia na reoreliéfe a jej modelovanie pomocou komplexného digitálneho modelu reliéfu. *Geografický časopis* 44:342–355 (in Slovak)
29. Gruter JW (1984) Radiation nomenclature. 2nd solar energy programme of the CEC, Project F, Solar radiation data. CEC, Brussels
30. Muneer T (1990) Solar radiation model for Europe. *Build Serv Eng Res Technol* 11:153–163
31. Brest CL, Goward SN (1987) Deriving surface albedo measurements from narrow band satellite data. *Intl J Remote Sens* 8:351–367
32. Dubayah R (1992) Estimating net solar radiation using Landsat Thematic Mapper and digital elevation data. *Water Resour Res* 28:2469–2484
33. Duguay CR, LeDrew EF (1992) Estimating surface reflectance and Albedo from LANDSAT-5 thematic mapper over rugged terrain. *Photogramm Eng Remote Sens* 57:551–558
34. Dubayah R, Loechel S (1997) Modeling topographic solar radiation using GOES data. *J Appl Meteorol* 36:141–154
35. Becker S (2001) Calculation of direct solar and diffuse radiation in Israel. *Int J Climatol* 21:1561–1576
36. Kitler R, Mikler J (1986) *Základy využívania slnečného žiarenia*. Bratislava, Veda (in Slovak)
37. Beyer HG, Costanzo C, Heinemann D (1996) Modifications of the Heliosat procedure for irradiance estimates from satellite images. *Sol Energy* 56:207–212
38. Rigollier Ch, Lefèvre M, Wald L (2001) Heliosat version 2, integration and exploitation of networked solar radiation databases for environment monitoring (SoDa), European commission project no. IST-1999-122245, report. http://www.helioclim.org/publications/heliosat2_d3.2.pdf. Accessed 14 December 2010
39. Kasten F, Czeplak G (1980) Solar and terrestrial radiation dependent on the amount and type of cloud. *Sol Energy* 24:177–189
40. Cano D, Monget JM, Albuisson M, Guillard H, Regas N, Wald L (1986) A method for the determination of the global solar radiation from meteorological satellite data. *Sol Energy* 37:31–39
41. Hofierka J, Šúri M (2002) The solar radiation model for Open source GIS: implementation and applications. In: Ciolli M, Zatelli P (eds) *Proceedings of the open source free software GIS – GRASS users conference*, Trento, Italy, 11–13 Sept 2002, CD-ROM
42. Šúri M, Huld TA, Dunlop ED, Ossenbrink HA (2007) Potential of solar electricity generation in the European Union member states and candidate countries. *Sol Energy* 81:1295–1305
43. Wald L (2006) Available databases, products and services. In: Dunlop ED, Wald L, Šúri M (eds) *Solar energy resources management for electricity generation from local level to global scale*. Nova Science Publishers, New York, pp 29–41
44. Hofierka J, Cebecauer T (2007) Priestorová a časová distribúcia slnečného žiarenia na georeliéfe Slovenska. In: Štrélcová K, Škvarenina J, Blaženec M (eds) *Bioclimatology and natural hazards*, CD-ROM proceedings of the international scientific conference, Polana nad Detvou, Slovak Bioclimatological Society, Technical University Zvolen (in Slovak), Slovakia, 17–20 Sept 2007
45. Zyl JJ (2001) The shuttle radar topography mission (SRTM): a breakthrough in remote sensing of topography. *Acta Astronaut* 48:559–565
46. Hutchinson MF, Booth TH, McMahon LP, Nin HA (1984) Estimating monthly mean values of daily total solar radiation for Australia. *Sol Energy* 32:277–290
47. Hulme M, Conway D, Jones PD, Jiang T, Barrow EM, Turney C (1995) A 1961–1990 climatology for Europe for climate

- change modelling and impact applications. *Int J Climatol* 15:1333–64
48. Zelenka A, Czeplak G, D'Agostino V, Josefson W, Maxwell E, Perez R (1992) Techniques for Supplementing Solar Radiation Network Data, Technical Report, International Energy Agency, # IEA-SHCP-9D-1. Swiss Meteorological Institute, Switzerland
 49. D'Agostino V, Zelenka A (1992) Supplementing solar radiation network data by co-kriging with satellite images. *Int J Climatol* 12:749–761
 50. Beyer HG, Czeplak G, Terzenbach U, Wald L (1997) Assessment of the method used to construct clearness index maps for the new European solar radiation atlas (ESRA). *Sol Energy* 61:389–397
 51. Hofierka J, Parajka P, Mitasova M, Mitas L (2002) Multivariate interpolation of precipitation using regularized spline with tension. *Trans GIS* 6:135–150
 52. Remund J, Wald L, Lefèvre M, Ranchin T, Page J (2003) World-wide linke turbidity information. In: Proceedings of ISES solar world congress on solar energy for a sustainable future, Göteborg, Sweden, 14–20 June 2003
 53. Gueymard CA (2008) Fixed or tracking solar collectors? Helping the decision process with the solar resource enhancement factor. In: Proceedings of optical modeling and measurements for solar energy systems II conference, SPIE, vol 7046, San Diego
 54. Šúri M, Huld TA, Dunlop ED, Hofierka J (2007) Solar resource modelling for energy applications. In: Peckham RJ, Jordan G (eds) Digital terrain modelling, development and applications in a policy support environment, Series lecture notes in geoinformation and cartography. Springer, Heidelberg, pp 259–273
 55. Ruiz-Arias JA, Cebecauer T, Tovar-Pescador J, Šúri M (2010) Spatial disaggregation of satellite-derived irradiance using a high-resolution digital elevation model. *Sol Energy* 84:1644–1657

Toxic Chemical Risks

EDWARD A. LAWS

Louisiana State University, School of the Coast and Environment, Baton Rouge, LA, USA

Article Outline

Glossary

Definition of the Subject

Introduction

Lifestyle-Associated Exposure to Toxic Chemicals

Exposure to Toxic Chemicals in the Home

Future Directions

Bibliography

Glossary

Alkaloid A colorless, complex, and bitter organic base containing nitrogen and usually oxygen that is often physiologically active.

Analgesic Painkiller.

Apoptosis A genetically directed process of cell self-destruction that eliminates DNA-damaged, superfluous, or unwanted cells.

Cold turkey Cold flashes with goose bumps, one of the symptoms associated with withdrawal from a drug to which a person has become addicted.

Dopamine A neurotransmitter in the brain that plays an important role, inter alia, in behavior and reward.

Opioid A chemical that binds to and activates the opioid receptors in the nervous system and gastrointestinal tract; includes both natural opiates (derived from opium) and synthetic drugs possessing narcotic properties similar to opiates but not derived from opium.

Secondhand smoke Smoke inhaled by a nonsmoker and consisting of mainstream smoke (smoke exhaled by a smoker) and sidestream smoke (smoke from the end of a lighted cigarette, pipe, or cigar).

Teratogenic Causing developmental malformations.

Definition of the Subject

At the present time there are approximately 100,000 chemicals in the environment, with an additional 500–1,000 added each year. Although not all these chemicals are toxic, merely keeping up-to-date information on the possible toxicity of so many compounds is a prodigious task. Federal agencies such as the Environmental Protection Agency (EPA), Food and Drug Administration (FDA), and Consumer Product Safety Commission (CPSC) are responsible for establishing rules and guidelines that ensure, insofar as possible, that risks associated with exposure to toxic chemicals are within acceptable limits. Legislation such as the Clean Air Act, originally passed in 1963, and Clean Water Act, a 1977 amendment to the 1972 Federal Water Pollution Control Act, have done much to advance the cause of environmental safety. Ironically perhaps, in the United States the primary sources of exposure to toxic chemicals and poisons are related to lifestyle issues, including in particular smoking, drinking (alcohol), use/abuse of pharmaceutical and

recreational drugs, and obesity. Most incidents of poisoning occur in the home, where unintentional exposure to toxic chemicals is often the result of ignorance and/or carelessness. Education and public awareness have therefore become important components of the effort to address the issue of exposure to toxic chemicals in the United States.

Introduction

On the first day of class students in an honors section of an introductory environmental science class at Louisiana State University were asked to answer a series of multiple-choice questions to determine how much they knew about the subject before taking the course. One of the questions asked (Dr. Vince Wilson, personal communication):

The statement, “Water can be a poison,” is an example of

1. No agent has a single effect
2. The water quality of the Baton Rouge Public Water System
3. Dose makes the poison
4. A sensitive population
5. Toxicology is an empirical science

The correct answer is “dose makes the poison,” the translation of a statement made in German by the sixteenth century physician Paracelsus. It underscores the fact that many substances commonly considered to be toxic are presumably harmless if the dose or degree of exposure is sufficiently small. This line of reasoning underlies the concept of an acceptable daily intake (ADI), for example, of food additives and contaminants [1, p. 2]. The ADI is a dose of the substance that would be expected to be associated with no appreciable adverse health effects if administered over the lifetime of a healthy person. On the other hand, consumption of too much of an otherwise benign substance can have serious health consequences. Water intoxication, for example, is a potentially fatal condition that occurs when consumption of too much water causes an abnormal balance of electrolytes in the body.

In judging the risk associated with exposure to toxic chemicals, one must consider the duration of exposure as well as the dose. In toxicology the distinction is typically made between short-term or acute exposure and long-

term or chronic exposure. Experimentally acute and chronic exposures are defined to be exposures lasting up to 14 days and greater than 1 year, respectively, with “intermediate exposure” used to characterize exposures lasting 14 days to 1 year [2]. These definitions are arbitrary, and in practice exposure times may vary continuously from seconds to an entire lifetime.

Alcohol consumption is a good example of an issue linked to exposure time. In most states in the United States, a person is considered to be intoxicated if his/her blood alcohol content exceeds 0.08%. Consumption of alcohol can have fatal consequences if the blood alcohol content approaches ~0.40%. Death results from alcohol’s depressant effect on vital areas of the brain that control consciousness, respiration, and heart rate. To provide some context to these numbers, a person weighing ~100 pounds (45 kg) who consumed the contents of two 12-ounce cans of beer (5% alcohol content) over a short timeframe (~30–40 min) would likely have a blood alcohol content of ~0.10% (i.e., above the legal limit for driving) and a blood alcohol content of ~0.40% (i.e., potentially lethal) if he/she consumed eight 12-ounce cans of beer over the same timeframe. On the other hand, a 150-pound (68 kg) person who consumed the same quantities of beer over the same timeframe would likely have blood alcohol levels of only ~0.06% (i.e., below the legal limit for driving) and 0.25% (intoxicated and perhaps experiencing serious side effects, but probably not life-threatening), respectively. Incidentally, the tendency of people to vomit when they have had too much to drink is a natural defense mechanism and allows the body to rid itself of alcohol that has not already been absorbed into the bloodstream. Unfortunately this defense mechanism is not infallible. When some people have had too much to drink they lose consciousness and never recover, either because of the effect of the alcohol on their brain or because they choke on their own vomit and are unable to clear their air passages.

The effects of consuming 12-ounce cans of beer are very different if the timeframe is longer. If a 100-pound person consumed one 12-ounce can of beer per day for 8 days, for example, his/her blood alcohol level would probably never rise to more than 0.05%, below the legal limit for driving and far below the level associated with potentially fatal effects.

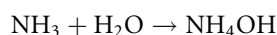
Perhaps surprisingly, most (~90%) exposures to toxic chemicals occur in the home and involve everyday household items. Not surprisingly, most exposures (85–90%) are accidents, although some are intentional (e.g., suicides). With the notable exception of cigarette smoke, about 75% of cases in the United States involve ingestion of a poison (as opposed, e.g., to inhalation of a toxin or exposure via dermal contact), and, with the exception of alcohol consumption, over half the cases involve children under the age of six. Interestingly, the greatest public health problems associated with exposure to toxic chemicals involve lifestyle decisions as opposed, for example, to living downwind or downstream from an obvious source of pollution.

Lifestyle-Associated Exposure to Toxic Chemicals

Smoking

In the United States and many other countries, by far the single greatest source of exposure to toxic chemicals is tobacco smoke, primarily the result of cigarette smoking. Cigarette smoke contains several hundred toxins, of which 44, the so-called Hoffmann analytes [3–5], are believed to be most relevant to smoking-related health effects [6]. These include the following:

- Ammonia (NH₃): Nicotine, which is the addictive component of cigarettes, exists in two forms, bound and free. The free form of nicotine vaporizes more readily, and from the vapor phase is rapidly absorbed by the lungs and transported by the bloodstream to the brain. Adding ammonia to cigarettes facilitates the conversion of bound to free nicotine and thereby enhances the pleasurable effects associated with secretion of dopamine by brain neurons that have been stimulated either directly or indirectly by nicotine. Ammonia is also a toxic gas that can damage the respiratory system when it combines with water to form ammonium hydroxide:



NH₄OH is highly alkaline and corrosive. Exposure to NH₄OH is associated with coughing, sore throat, shortness of breath, and a burning sensation, all of which may be associated with smoking.

- Arsenic (As): Arsenic is a well-known toxin that has been used in pesticides for more than 1,000 years.

The most commonly used formulation has been lead arsenate, which was used extensively for control of insect pests until the end of World War II, when more effective synthetic organic insecticides became available. However, it is estimated that smokers breathe in between 0.8 and 2.4 µg of As per pack of cigarettes [7] due to continued use of arsenic-containing pesticides by tobacco growers. The toxicity of As is related to its disruptive effects on a wide variety of metabolic processes leading to, *inter alia*, oxidative stress, lactic acidosis, anemia, central nervous system dysfunction, and apparent thiamine deficiency.

- Benzene (C₆H₆): Benzene, a human carcinogen and known cause of leukemia, is found in cigarette smoke, which is believed to account for about half of human exposure. Chronic exposure to benzene can result in a variety of health problems including anemia, leukemia, and adverse effects on the immune system [8, 9].
- Benzo[a]pyrene (C₂₀H₁₂): Benzo[a]pyrene is a five-ring polycyclic aromatic hydrocarbon formed by the fusion of a benzene ring with pyrene. It is found in all smoke associated with the combustion of organic matter, including tobacco leaves. Metabolites of benzo[a]pyrene have been shown to be mutagenic and are highly carcinogenic [10]. Amounts of benzo[a]pyrene in cigarettes range from ~3 to 30 ng per cigarette and are roughly in proportion to declared tar values [11]. The carcinogenic effects of benzo[a]pyrene are associated with binding of its diol epoxide metabolites to the tumor suppressor protein P53, which normally functions as a cell cycle regulator. Damage to P53 is related to more than half of all human cancers and to an even higher percentage of lung cancers [12, 13].
- Cadmium (Cd): In general leafy vegetables such as lettuce and spinach contain high levels of cadmium, which they accumulate from the soil. This is a natural phenomenon, and tobacco leaves are no exception. A cigarette smoker will typically take in about 3.0 µg of Cd per pack, of which 25% will be absorbed [14]. Chronic exposure to Cd is associated with kidney and liver damage, and in severe cases to softening of the bones [15]. Smokers typically have about twice the body burden of Cd as nonsmokers [16, 17].

- **Carbon monoxide (CO):** Carbon monoxide is a colorless, odorless gas that results from the combustion of organic matter with inadequate amounts of oxygen. The concentration of CO in the air inhaled with cigarette smoke is about 2%, of which ~85% is absorbed [18, 19]. A person smoking 20 cigarettes per day (the minimum number of cigarettes per pack in the United States is 20) would be expected to absorb ~200 mg of CO [20], considerably higher than the maximum of 24 mg d⁻¹ that can be safely tolerated [18]. The adverse effects of CO exposure are typically associated with its tendency to bind to hemoglobin and hence to interfere with the ability of hemoglobin to absorb and transport oxygen. Ironically, the affinity between CO and hemoglobin is about 230 times greater than the affinity between O₂ and hemoglobin [21, 22]. Additional adverse effects may be caused by the tendency of CO to bind to myoglobin, a protein that binds iron and oxygen in muscle tissues, and to cytochrome oxidase, an enzyme in the respiratory electron transport chain.
- **Formaldehyde (CH₂O):** Formaldehyde is a colorless but pungent gas found in cigarette smoke as a result of the pyrolysis of saccharides used as tobacco ingredients [23–25]. It is carcinogenic [26–29] and causes damage to the lungs when inhaled [30, 31]. The amount of formaldehyde inhaled in cigarette smoke is closely correlated with the tar content of the cigarette and ranges as high as 40–50 µg per cigarette [20]. Estimated exposure levels in the oral cavity during smoking range as high as 115 parts per million (ppm), which is considerably higher than the Occupational Safety and Health Administration (OSHA) short-term (15 min) exposure limit (STEL) of 2 ppm.
- **Hydrogen cyanide (HCN):** Hydrogen cyanide is a colorless, poisonous gas that was used as a genocidal agent by the Nazis during World War II. A byproduct of burning tobacco, HCN is found in cigarette smoke in amounts sufficient to produce exposures of 144–351 µg per cigarette [6, 32]. To put this number in context, the STEL for HCN is 4.7 ppm in the air one breathes. An average person inhales about 11 m³ of air per day or 115 L in 15 min. If the air contained 4.7 ppm HCN vapor, the person would be inhaling about 0.54 mL or 22 µmol of HCN vapor in 15 min (this calculation assumes that HCN vapor behaves like an ideal gas and that the temperature is 25 °C or 77 °F). Twenty-two micromoles of HCN have a mass of 594 µg. Thus smoking 2–4 cigarettes in 15 min would very likely lead to the inhalation of a quantity of HCN greater than the OSHA short-term exposure limit for this toxic gas. Headache, nausea, and vertigo are symptoms associated with chronic exposure to HCN [33].
- **Nicotine (C₁₀H₁₄N₂):** Nicotine is an alkaloid that is responsible for the addictive nature of tobacco use. However, it has also been implicated in DNA damage, oxidative stress, and inhibition of apoptosis [34, 35]. The latter effect could be a factor in the etiology of smoking-associated cancer of the lungs and oral cavity, that is, inhibition of programmed cell death may allow the multiplication of “chemically initiated cells” [35]. Doses of nicotine associated with cigarette smoking are highly correlated with the tar content of cigarettes and range as high as 1.0–1.2 mg per cigarette [20].
- **TSNAs:** This acronym stands for tobacco-specific N-nitrosamines, a category that includes four compounds linked to smoking-related cancers: (1) N'-nitrosoanatabine (NNA), (2) 4-methyl-N-nitrosamino-1-(3-pyridyl)-1-butanone (NNK), (3) N-nitrosoanatabine (NAT), and (4) N-nitrosoanabasine (NAB). Together with the polycyclic aromatic hydrocarbons (e.g., benzo[a]pyrene) these are among the most potent cancer-causing agents in cigarette smoke [36–38]. NNA, NNK, and NNAL, for example, cause cancer in mice, rats, and hamsters [36]. The TSNAs are produced during tobacco processing and smoking and are derived from nicotine and minor tobacco alkaloids. A variety of experimental studies have demonstrated that TSNAs are in fact procarcinogens, that is, agents that require metabolic activation before they become carcinogenic. Their activation involves α-hydroxylation to form unstable α-hydroxynitrosamines, which then decompose to diazohydroxides that react with cellular components, including DNA and hemoglobin. There is good evidence that TSNAs contribute to the increased risk for cancer (1) of the upper digestive tract among tobacco chewers and (2) of lung cancer, especially pulmonary adenocarcinoma, in smokers [36].

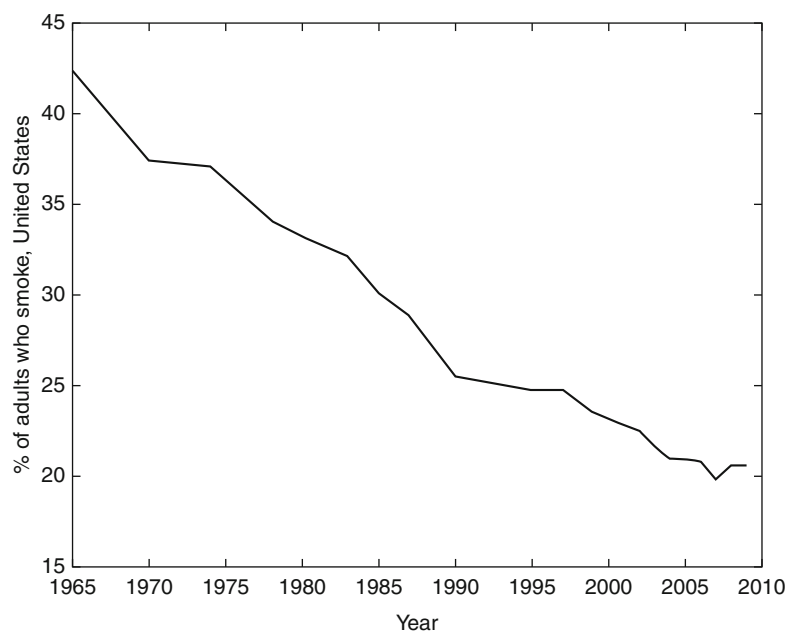
Considering the large number of toxic compounds in cigarette smoke, including in particular those implicated in causing cancer, it should come as no surprise that smoking is associated with a considerable health risk. Approximately 440,000 people die each year in the United States as a result of cigarette smoking, including approximately 50,000 who die from exposure to secondhand smoke. The combined death toll accounts for almost 20% of all deaths in the United States annually (the death rate in the United States is about 2.4 million people per year) and exceeds all deaths associated with alcohol, AIDS, car accidents, illegal drugs, murders, and suicides combined [39]. The lifespan of smokers is about 14 years shorter than that of nonsmokers [40]. Statistics compiled by the Centers for Disease Control (CDC) indicate that persons who smoke increase their risk of dying from bronchitis and emphysema roughly tenfold. Men and women who smoke increase their risk of dying from lung cancer by more than a factor of 22 and by nearly a factor of 12, respectively [39]. In the United States, cigarette smoking is estimated to be responsible for \$193 billion in annual health-related economic losses

(this figure includes \$96 billion in direct medical costs and another ~\$97 billion in lost productivity), a figure that translates to about \$10.47 in economic costs per pack of cigarettes sold [41].

The good news is that smoking in the United States is on the decline. The percentage of adults who smoke has declined more-or-less monotonically from a high of 42.4% in 1965 to 20.6% in 2008–2009 (Fig. 1). The percentage of high school students who smoke has declined even more rapidly, from a high of 36% in 1997 to 19.5% in 2009 [42].

The pattern evident in Fig. 1 reflects increasing public awareness of the health risks associated with cigarette smoking and, in recent years, an effort by the federal government to reduce tobacco-related diseases and deaths. The National Tobacco Control Program (NTCP), created by the CDC in 1999, provides funding and technical support to state and territorial health departments with the following goals:

- Prevent initiation of smoking by youth.
- Promote quitting among both adults and youth.



Toxic Chemical Risks. Figure 1

Percentage of adults in the United States who smoke [42]

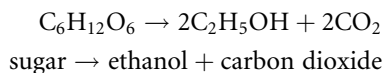
- Eliminate exposure to secondhand smoke.
- Identify and eliminate disparities among population groups.

More information about the program is available at the NTCP Web site:

http://www.cdc.gov/tobacco/tobacco_control_programs/ntcp/index.htm

Alcohol Consumption

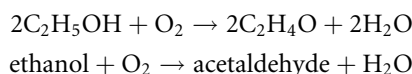
Alcohol is probably the oldest drug used by human beings, with commercial production of wine traceable to as early as 1,500 B.C. [43]. The alcohol in question is ethanol (ethyl alcohol), which is produced by the fermentation of sugar via the overall reaction:



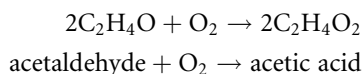
The pleasurable effects of ethanol are associated with its effects on the brain. From the digestive tract ethanol rapidly passes to the blood stream and from there to all parts of the body, including the brain. Ethanol depresses the membrane responsiveness of all neurons in the body, which explains the slower reaction time and sedative effects associated with ethanol consumption. Ethanol also helps to increase the release of dopamine by brain neurons, which results in artificial feelings of pleasure. This interference with normal brain functionality eventually causes the brain to reduce dopamine activity, leading over time to a greatly reduced ability to experience pleasure by normal means. This in turn leads to alcohol addiction.

In addition to being addictive, ethanol is also toxic. In fact, it is routinely used as an antiseptic because it kills bacterial cells. At a sufficiently high concentration it will kill virtually any cell. Even doses typical of recreational consumption of alcoholic beverages lead to a loss of a small number of cells in the brain and liver. If repeated often enough, this can lead to health problems, for example, hepatic cirrhosis and cancers. The fact that the ethanol content of wine is about 10%, for example, is no accident. The yeasts that typically mediate the fermentation process in winemaking die if the ethanol content of the product exceeds about 12%.

The organ primarily responsible for eliminating ethanol from the human body is the liver, where ethanol is converted to acetaldehyde via the enzyme alcohol dehydrogenase:



This does not solve the toxicity problem, however, because acetaldehyde itself is toxic and is responsible for many of the symptoms associated with hangover. In the liver, acetaldehyde is converted to acetic acid ($\text{C}_2\text{H}_4\text{O}_2$) via acetaldehyde dehydrogenase. Simplistically:



Problem solved? The answer is yes if a person consumes ethanol infrequently enough, but as noted in the introduction to this chapter, the dose makes the poison. Serious health problems can arise if a person consumes ethanol too frequently over a long period of time. Health problems associated with such long-term ethanol consumption include the following [44]:

- Liver damage: Alcoholic hepatitis, cirrhosis of the liver, and cancer of the liver can all result from chronic ethanol consumption. Alcoholic hepatitis is an inflammation of the liver caused by excessive ethanol consumption. Cirrhosis of the liver is caused by a gradual deterioration of the functionality of the liver and is characterized by replacement of normal liver tissue by scar tissue.
- Brain damage: There are a variety of symptoms here, including impaired judgment, vision, and coordination (leading in some cases to accidents), cell death (leading to dementia and in some cases to paralysis), and in cases of acute overdoses, coma and death.
- Impacts on the reproductive system: Here again, there are many symptoms, including impotence due to hormone imbalances, birth defects caused by mutations of egg or sperm cells, fetal alcohol syndrome (damage to the fetus associated with the fact that ethanol passes from the mother's bloodstream across the placenta to the fetus), teratogenic effects, and breast cancer.

- Damage to the digestive tract: Impacts include inflammation and cancer of the throat, esophagus, stomach, pancreas, small intestine, and colon.

The greatest risks from ethanol consumption are associated with the development of a condition called alcoholism. Simplistically, this is a condition that develops when a person's brain has been sufficiently altered by ethanol consumption that the person must drink to feel normal. The result is compulsive drinking. Failure to drink leads to serious withdrawal symptoms that include nausea, sweating, shakiness, anxiety, and in a small percentage of cases, to a condition called delirium tremens, which can be fatal. Most people who become alcoholics and subsequently stop drinking do so only with assistance.

How many people are alcoholics? About 10% of Americans are estimated to have "an alcohol problem" and 7.7% "abuse alcohol or are alcoholic" [45]. Thirty-eight percent of American families report a "history of alcoholism" [46].

The costs associated with alcoholism in terms of lives, human health, and societal impacts are very large. In the United States, more than 100,000 people die annually as a result of excessive alcohol consumption. This is about 4% of the annual death rate in the United States. The specific causes of death include, inter alia, drunken driving, cirrhosis of the liver, cancer, and stroke. In the case of drunken driving, many victims are not themselves alcoholics but instead are innocent bystanders. Roughly half of all homicides, 40% of assaults, and one-third of all fatal car accidents, suicides, and hospital admissions in the United States are alcohol related. The economic costs associated with alcohol abuse and alcoholism in the United States are roughly \$230 billion per year, almost 80% more than the costs related to all other addictive drugs combined [45, 47].

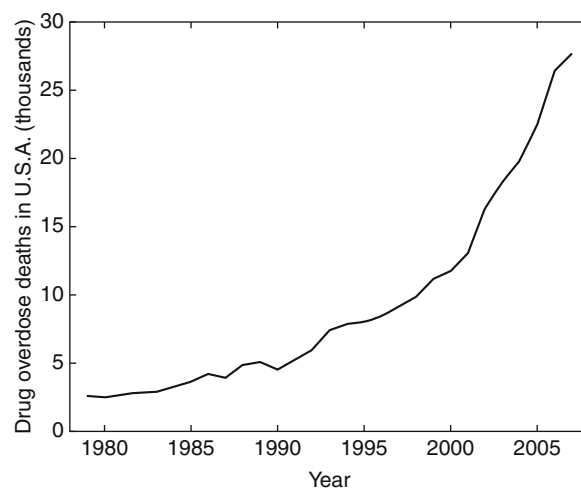
Most people who drink are not alcoholics. In the United States, roughly 44% of adults (18 years and older) consume an average of at least one drink per month, including 60% of persons 21–34 years of age, 46% for adults aged 60–64, and 33% of persons 65 or older [48]. Clearly only a small fraction of persons who consume ethanol become alcoholics. The risk of becoming an alcoholic is to some extent a function of genetics. Alcoholism runs in families [49, 50]. But the

risk of becoming an alcoholic is also a function of one's environment, including peer pressure, the influence of family and friends, and the ease with which one can obtain alcohol [48].

Drug Abuse/Misuse

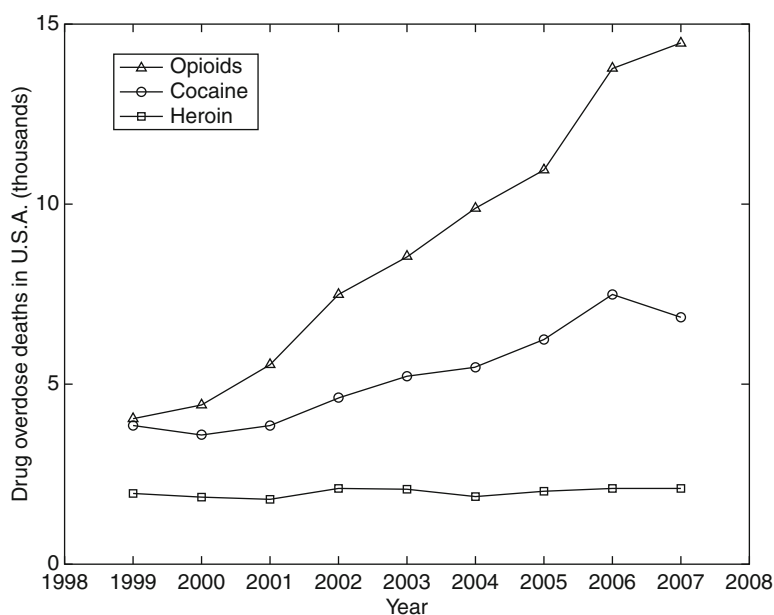
In the last 30 years, the number of deaths associated with unintentional drug overdoses has increased dramatically in the United States (Fig. 2). Most of the increase has not been due to the use of illegal drugs such as cocaine and heroin but rather due to misuse/abuse of prescription painkillers (analgesics) that are opioids (Fig. 3). Opiates are compounds derived from opium poppies and include morphine and codeine. Opioid analgesics are synthetic or semisynthetic analogues of opiates and include oxycodone (e.g., OxyContin®), hydrocodone (e.g., Vicodin®), propoxyphene (e.g., Darvon®), and hydromorphone (e.g., Dilaudid®) [51].

Naively one might assume that deaths associated with opioid analgesics are primarily linked to little children getting into medicine cabinets or old people mixing up their pills, but that is not the case. According to Paulozzi [53], the trend in opioid-related deaths apparent in Fig. 3 is due to the increasing use of "prescription drugs, especially opioid painkillers,



Toxic Chemical Risks. Figure 2

Number of deaths per year in the United States associated with unintentional drug overdoses [52]



Toxic Chemical Risks. Figure 3

Number of deaths per year in the United States associated with unintentional overdoses of opioids, cocaine, and heroin [54, 55]

among people during the working years of life.” A high percentage of the people who die have a history of drug abuse, and many have no prescription for the drug. Mixing of prescription drugs with illegal drugs is common, and “some alter the prescription drugs by crushing and snorting them or dissolving and injecting them” [53].

The problems associated with the use of opioid analgesics are an excellent example of the adage that the dose makes the poison. When used as prescribed, opioid analgesics can be very effective painkillers. They bind to specific proteins (opioid receptors) in the brain, spinal cord, and gastrointestinal tract, and in this way they can block the transmission of pain signals to the brain.

There are side effects associated with opioid/opiate use. One is suppression of coughing, which is why codeine, for example, is incorporated into a number of cough syrups and cold remedies. Another is constipation, which explains why loperamide, an opioid derived from piperidine, is used to treat diarrhea. More insidious side effects include drowsiness and a feeling of euphoria. The former is associated with the lethal effects of an opioid overdose, which leads to

a suppression of breathing and may ultimately cause respiratory failure. The latter effect is responsible for the abuse and compulsive use of opioids. Persons addicted to opioids experience withdrawal symptoms if use of the drugs is reduced or stopped. These symptoms include muscle and bone pain, insomnia, diarrhea, vomiting, involuntary leg movements, and cold flashes with goose bumps (i.e., cold turkey). Interestingly, the synthetic opioid methadone blocks the effects of other opioids (including heroin), eliminates withdrawal symptoms, relieves drug craving, and has been used for decades to treat opioid addicts [56]. Fortunately, properly supervised use of opioid analgesics seldom results in clinical addiction.

Cocaine is an illegal drug and one of the most addictive of all stimulants. Derived from the coca bush, cocaine is marketed “on the street” in two forms: a water-soluble hydrochloride salt and a water-insoluble cocaine base. A cocaine user can either inject the water-soluble form with a hypodermic needle or snort (i.e., inhale) the hydrochloride salt powder through the nose, where it is readily absorbed through nasal tissues into the bloodstream. The base form is processed to produce a substance known as crack that

can be smoked. The term “crack” comes from the crackling sound heard when the mixture is smoked. Once it has entered the blood stream, cocaine travels to the brain, where it produces a euphoric effect that is responsible for its popularity and addictive properties. In a nutshell, cocaine causes dopamine to accumulate at nerve synapses (small gaps between two neurons) in the brain and thereby heightens the pleasurable sensations associated with dopamine secretion. Rather than stimulating the secretion of dopamine by a neuron, the mechanism involves blockage of the natural process by which dopamine is recycled from the synapse to the neuron by specialized proteins called dopamine transporters.

The pleasurable effects of cocaine appear rapidly but are of short duration. When cocaine is absorbed rapidly via either injection or smoking, the effects are intense but dissipate after 5–10 min. Snorting is associated with a delayed response that is less intense but persists for 15–30 min [57].

In addition to the general euphoric sensation associated with cocaine use, there are other psychological effects. Users feel mentally alert, talkative, and energetic and temporarily may feel less need for sleeping and eating. Physiological effects include constricted blood vessels and an increase in body temperature, pulse, and blood pressure. Medical complications arising from the physiological effects of cocaine include the following [57]:

- Cardiovascular effects: arrhythmias (heart rhythm disturbances), heart attacks
- Neurological effects: strokes, seizures, headaches, coma
- Gastrointestinal complications: abdominal pain, nausea

Deaths associated with cocaine use are typically associated respiratory failures preceded by cardiac arrest or seizures. In addition, there is an even greater danger associated with the use of cocaine during alcohol consumption. Alcohol and cocaine together are the most common two-drug combination responsible for drug-related deaths [57].

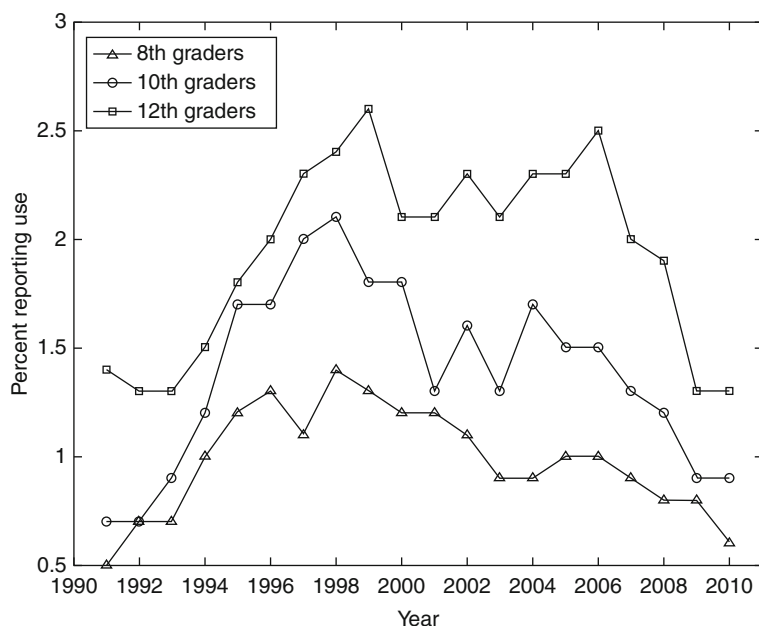
Clearly there is a very significant risk associated with cocaine abuse, and with the exception of its very limited use as a topical anesthetic in nasal and lacrimal duct surgery, it is associated with no beneficial

applications. Moreover, there are presently no medications approved by the Food and Drug Administration (FDA) to treat cocaine addiction. In 2007, cocaine abuse accounted for 13% of all admissions to drug abuse treatment programs in the United States, and a majority of those cocaine addicts were abusing multiple drugs [57]. Treatment of cocaine addiction is a complex process, not only because of the frequent problem of multiple-drug abuse but also because of the complex biological (e.g., changes in the brain), social, familial, and environmental problems that must typically be addressed in the treatment process. Flynn et al. [58] examined the factors that contributed to long-term recovery from cocaine dependence of 708 patients from 45 treatment programs in the United States and concluded that only 33% of the sample were “highly favorable” at follow-up. The major reasons cited for improvements were “motivations to change, positive influences of family, strength from religion and spirituality, and help from drug treatment.”

Given the highly addictive nature of cocaine, the absence of approved medications for treatment of addiction, and the mixed success of treatment programs, Benjamin Franklin’s observation that “an ounce of prevention is worth a pound of cure” applies (Franklin’s comment was advice related to fighting fires; he founded Philadelphia’s Union Fire Company). The good news is that use of cocaine by teenagers has been declining in the United States since roughly 1997 (Fig. 4). This decline, however, has followed a period of increase during the 1990s. It is apparent from Fig. 4 that the percentage of US teenagers who use cocaine is little different now than in 1990.

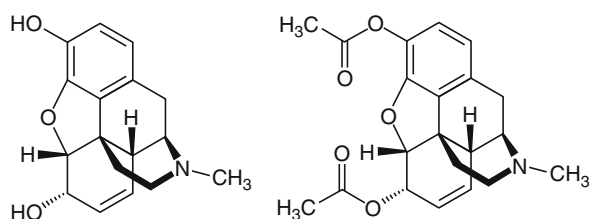
Heroin is the most abused and the most rapidly acting of the natural opiates. It is obtained by the acetylation of the two hydroxyl groups of morphine with acetyl chloride, that is, the two –OH groups are replaced with –OCOCH₃ groups (Fig. 5). This modification makes heroin much less soluble in water than morphine but allows it to pass the blood-brain barrier much more rapidly than morphine. Interestingly, once inside the brain, heroin is converted to morphine and then rapidly binds to opioid receptors [60], leading to a surge of pleasurable sensation known as a “rush.”

Heroin may be injected, sniffed/snorted, or smoked. Intravenous injection is very common because



Toxic Chemical Risks. Figure 4

Trends in a 30-day prevalence of cocaine abuse among 8th, 10th, and 12th graders [59]



Toxic Chemical Risks. Figure 5

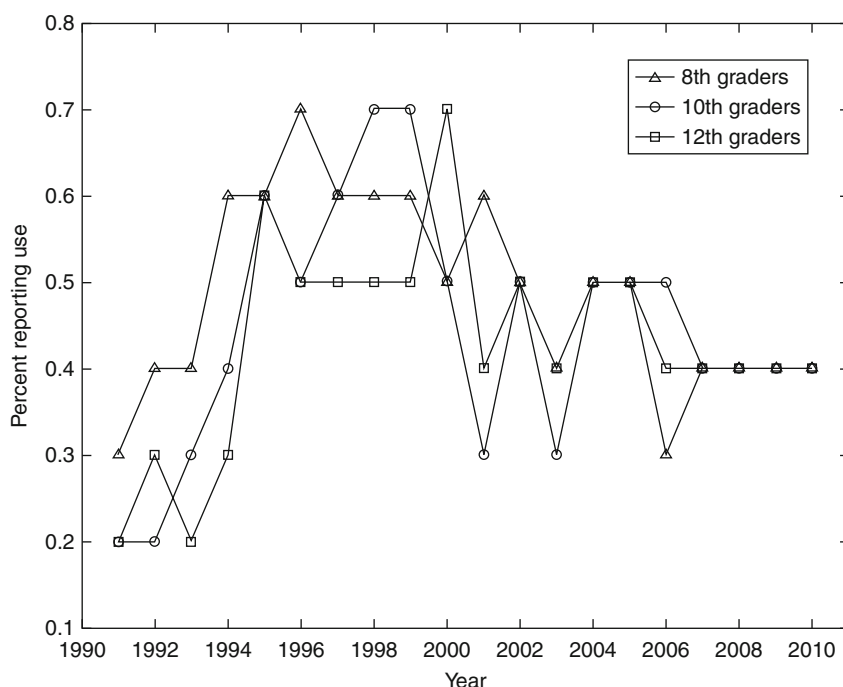
Structures of morphine (left) [61] and heroin (right) [62]

it provides the most rapid (less than 10 s) onset of euphoria. It is not uncommon for a heroin abuser to inject up to four times per day. Feelings of euphoria follow intramuscular injection and sniffing/snorting by 5–8 and 10–15 min, respectively [60]. Heroin is addictive regardless of how it is administered.

The risks associated with heroin addiction are numerous. For persons who inject, sharing of needles can lead to some of the most serious consequences: infection with hepatitis B and C, HIV, and other blood-borne viruses. Thirty-six percent of AIDS cases and an estimated 70–80% of the 35,000 new hepatitis C infections that occur in the United States each year are attributable to injection drug use [62, 63].

Then there are the effects of the heroin itself. After the initial rush, heroin users typically feel drowsy for several hours. Cardiac function and breathing slow, and in cases of overdose may slow to the point of death. On the street heroin is typically “cut” with other drugs or with substances such as sugar, starch, or powdered milk. As a result, users of street heroin are often unsure of the purity of the drug they are using. This uncertainty can easily lead to overdoses.

Chronic heroin injection can lead to a variety of medical problems, including scarred and/or collapsed veins, bacterial infections, and liver or kidney disease [62]. In addition to these physiological effects, heroin addicts gradually spend more and more of their time and energy obtaining and using heroin. “Once they are addicted, the heroin abuser’s primary purpose in life becomes seeking and using drugs” [62]. Withdrawal symptoms include restlessness, muscle and bone pain, insomnia, diarrhea, vomiting, involuntary leg movements, and the aforementioned cold turkey effects. These typically begin to manifest themselves within 1–2 days after the last dose of heroin and subside within 1 week, although in some cases they last for many months [62].



Toxic Chemical Risks. Figure 6

Trends in 30-day prevalence of heroin abuse among 8th, 10th, and 12th graders [59]

The good news is that there are a variety of effective treatments available for heroin addiction. As already noted, methadone effectively blocks the effects of heroin and has been used successfully to treat heroin addicts for decades. When properly used methadone is not intoxicating or sedating, it may be used safely for as long as 10 years or more. It is taken orally and suppresses withdrawal symptoms for 24–36 h. Very importantly, it relieves the craving for heroin, which is a major cause of relapse. A more recent development has been the use of buprenorphine, a semisynthetic opioid, for the same purpose. Use of buprenorphine is associated with less risk of addiction and can be prescribed in the privacy of a doctor's office. Another treatment, a combination of buprenorphine and naloxone (Suboxone®), is designed to minimize the possibility of abuse [62].

According to a 2003 survey of drug use in the United States [64], roughly 3.7 million people had used heroin at least once in their lives, 119,000 had used heroin within the previous

month, and 314,000 had used heroin within the previous year. The number of deaths attributable to heroin use in the United States has been constant for the last 20 years (Fig. 3) and has averaged about 2,000 per year. The fraction of teenagers who abuse heroin has been constant since ~2000 (Fig. 6) and is higher now than in the early 1990s. Thus heroin abuse and addiction continue to be serious problems in the United States.

Obesity

Food is perhaps the quintessential example of a substance for which the dose makes the poison. Most people do not think of food as a poison, but obesity is a serious problem in the United States and a disease formally recognized by the Centers for Disease Control and Prevention.

For a number of years public health agencies have used a metric called the body mass index (BMI) to determine whether a person's weight is normal or not.

If a person's weight and height are quantified in pounds and inches, respectively, the BMI is determined from the following equation:

$$\text{BMI} = 703 \frac{\text{weight}}{\text{height}^2}$$

Thus, for example, a person whose height is 70.3 in. and whose weight is 140.6 pounds would have a BMI of 20. As a rule of thumb, a person is categorized as follows based on his/her BMI:

- Underweight: <18.5
- Normal: 18.5–25.0
- Overweight: 25.0–30.0
- Obese: 30.0–40.0
- Extremely obese: >40.0

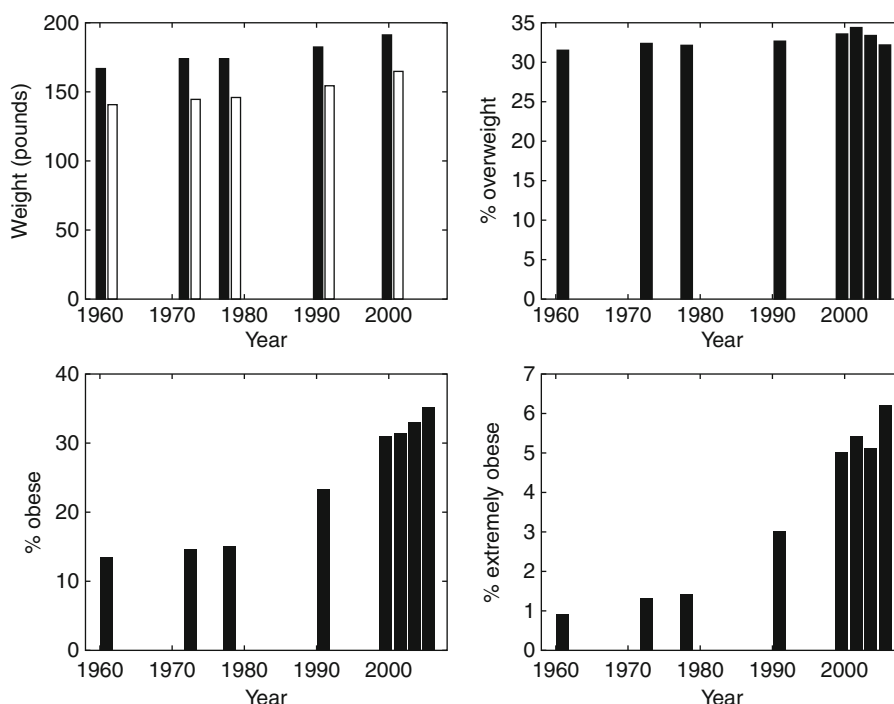
These are somewhat arbitrary definitions and in individual cases should be interpreted with the understanding that what is normal for one person

may not be normal for another. With that caveat, the BMI index has proven very useful for assessing the general health of a population relative to the issue of obesity.

The average weight of adults in the United States has been slowly increasing since 1960 (Fig. 7, upper left). The average weight of both men and women increased by ~25 pounds between 1960 and 2000. Interestingly, there was no change in the percentage of adults in the overweight category (Fig. 7, upper right), but there were rather dramatic increases in the percentages of adults characterized as obese (Fig. 7, lower left) and extremely obese (Fig. 7, lower right).

At the present time roughly 100 million adults in the United States are overweight or obese. Obesity substantially increases a person's risk of the following [67]:

- Morbidity from hypertension
- Dyslipidemia (abnormal amounts of lipids such as cholesterol and fat in the blood)



Toxic Chemical Risks. Figure 7

Metrics of US adults 20–74 years old with respect to weight. *Upper left:* Average weight of adult males (■) and females (□). *Upper right:* Percentage of adults with BMIs of 25–30. *Lower left:* Percentage of adults with BMIs of 30–40. *Lower right:* Percentage of adults with BMIs greater than 40 [65, 66]

- Type II diabetes (adult-onset diabetes)
- Coronary heart disease
- Stroke
- Gallbladder disease
- Osteoarthritis (degenerative joint disease)
- Sleep apnea (abnormal pauses or abnormally low breathing during sleep)
- Respiratory problems
- Endometrial (lining of the uterus), breast, prostate, and colon cancers
- Social stigmatization and discrimination

While there is no question that obesity significantly increases a person's risk of developing a number of potentially serious health problems, there is less agreement about what to do about obesity. The number of diet/exercise books in a typical bookstore is a good indication of the uncertainty associated with treatment of this disease. Guidelines recommended by the National Heart, Lung, and Blood Institute include the following:

- Goals
 - Initial target of weight loss should be 10%. If successful, further weight loss may be attempted.
 - Initial rate of weight loss should be 1–2 pounds per week for up to 6 months.
- Methods
 - Dietary therapy
 - Low-calorie diets with reduced fat content.
 - Reducing fat without reducing calories is futile.
 - Target caloric reduction should be 500–1,000 kcal per day.
 - Physical activity
 - The target here is at least 30 min of moderate-intensity physical activity on most, and preferably all, days of the week. Although physical activity will not be the primary determinant of weight loss, it will increase cardiorespiratory fitness and may decrease abdominal fat and help with maintenance of weight loss.
 - Behavior therapy
 - This translates to identifying ways to motivate the person who is trying to lose weight.

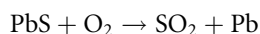
- Pharmacotherapy
 - This is the use of FDA-approved drugs and is only recommended if their use is supervised and their efficacy assessed by a physician.
- Weight loss surgery
 - This is only recommended in cases of clinically severe obesity when less invasive methods have failed and the patient is at high risk for obesity-associated morbidity or mortality.

Finally, there is the issue of weight loss maintenance. As with other lifestyle-related risks, the tendency to become obese can be a chronic problem that requires a lifetime of vigilance to overcome. The most effective strategy is a combination of dietary therapy, physical activity, and behavior therapy [67].

Exposure to Toxic Chemicals in the Home

Lead

Of all the toxic chemicals to which humans are exposed, lead has perhaps the longest history of exposure and today impacts the health of the greatest number of people. Lead is mined primarily from deposits of the mineral galena, or lead sulfide (PbS). Since metallic lead can be separated from PbS by heating at low temperatures easily achieved by burning wood or charcoal, it was not difficult for early civilizations to extract lead. The overall reaction is:



From written and archeological evidence, it is known that lead was used by the Egyptians as long ago as 1500 BC. The Romans used lead extensively to line their aqueducts and water mains, and both the Greeks and Romans used lead to line cooking vessels, since bronze pots tend to give food a bitter taste [68]. At least one author has suggested that endemic lead poisoning caused by the consumption of contaminated food and drink contributed significantly to the fall of the Roman Empire [69]. The use of lead or lead salts to sweeten wine was common in Europe after the ninth century, and associated outbreaks of lead poisoning were not infrequent. In colonial America lead poisoning was also a problem, mainly due to the use of lead

condensing tubes for distilling rum and the use of earthenware with a high lead content [68].

For obvious reasons, the principal uses of lead today are of an industrial rather than culinary nature. By far the major use, accounting for 87% of total lead consumption in the United States, is the production of batteries. However, lead batteries have never been a major source of human exposure to lead. During much of the twentieth century by far the most important source of human exposure was the use of lead in gasoline. Although leaded gasoline is no longer sold in the United States, lead was used extensively in the United States as a gasoline additive for roughly 60 years. This practice dates back to 1923, when tetraethyl lead was first introduced as an antiknock additive. This use was suspended during part of 1925 and 1926 pending the establishment of safety standards for its manufacture and handling, but after the spring of 1926 lead additives were commonly used in most gasoline sold in the United States and throughout the world. Beginning in 1960, organic alkyl lead compounds were blended into gasoline along with tetraethyl lead to further improve antiknock characteristics, and the additives were subsequently referred to as lead alkyls rather than just tetraethyl lead [68]. As of 1973, the lead content of gasoline sold in the United States averaged about 0.58 g/L.

Because of concern over pollution of the environment with lead, in 1974 the Environmental Protection Agency (EPA) requested a phasedown in the use of lead alkyls in gasoline to 0.45 g/L on January 1, 1975, and to 0.13 g/L as of January 1, 1979. In December 1974, however, the US Circuit Court of Appeals in Washington ruled against the EPA, stating that evidence did not support the EPA contention that automobile emissions contributed significantly to blood lead levels, and that the EPA regulation was arbitrary and capricious.

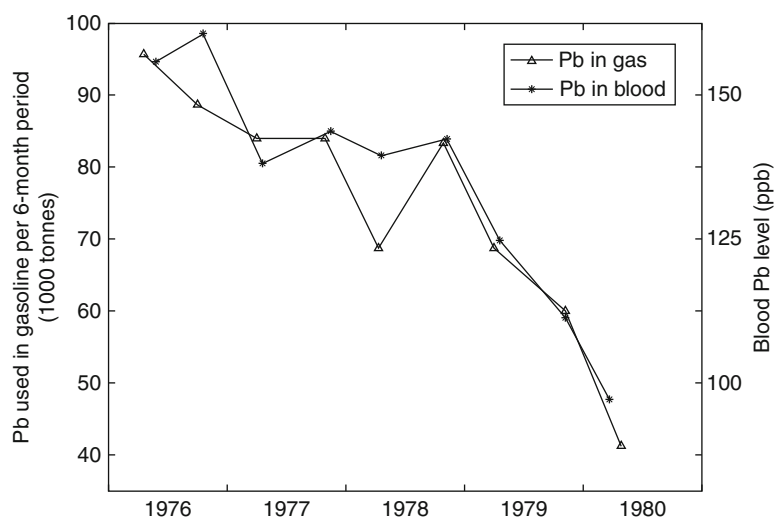
Since that time the legal status of gasoline lead additives has changed significantly. This change has resulted partly from the requirement for exhaust emission control devices on automobiles. The devices function poorly if at all when lead is present in the exhaust fumes, since the lead poisons the catalyst, which is designed to help oxidize unburned hydrocarbons. Furthermore, there is increasing awareness of the danger of environmental lead pollution due to the use of leaded

gasoline. Since July 1, 1977, all new cars sold in the United States have been required to run on unleaded gas, defined as gasoline containing less than 0.013 g/L lead. Finally, the government required that the lead content of leaded gasoline be reduced to 0.13 g/L by July 1, 1985, and to 0.10 g/L by January 1, 1986. The result of these regulations was a tenfold decline in the use of lead in gasoline in the United States between the early 1970s and 1986. Leaded gasoline is now completely phased out in the United States. European countries began slowly phasing out leaded gasoline in the 1980s and mandated elimination in the 1990s. Leaded gasoline was phased out in sub-Saharan Africa by 2006 [70]. China banned the sale of leaded gasoline in 2000, although production by clandestine factories for use in older vehicles continued until at least 2004 [71].

The decline in the blood lead levels of people in the United States during the phaseout of leaded gasoline (Fig. 8) is provocative considering the Circuit Court of Appeals 1974 ruling that evidence did not support the EPA's contention that automobile emissions contributed significantly to blood lead levels. During the period 1976–1980, blood lead levels declined in almost direct proportion to the use of leaded gasoline, suggesting that automobile emissions not only were contributing significantly to blood lead levels, but were in fact the dominant contributors.

Why all the concern about lead? With respect to toxicology, lead is a general metabolic poison. The pathological effects are varied, but in general reflect the tendency of lead to interact with proteins and hence to damage tissue and interfere with the proper functioning of enzymes [68]. Lead is known to inhibit active transport mechanisms involving ATP, to depress the activity of the enzyme cholinesterase, to suppress cellular oxidation–reduction reactions, and to inhibit protein synthesis [68]. Lead poisoning is also associated with the following problems:

- **Anemia:** Lead is known to disrupt several enzymes involved in the production of heme, a constituent of hemoglobin and various other respiratory pigments, and has been shown to interfere with the uptake of iron by red blood cells. Anemia associated with lead poisoning is undoubtedly caused in part by these effects. There is also evidence, however,



Toxic Chemical Risks. Figure 8

Change in blood lead levels during phaseout of leaded gasoline in the United States [72]

that lead causes a shortening of the life of red blood cells, probably due to disruption of the red cell membrane [68].

- **Damage to the Central Nervous System (CNS):** Damage to both the peripheral and central nervous system, including the brain, may be caused by a degeneration of nerve fibers and interference with the permeability of capillaries in the brain due to lead intoxication. The exact mechanism responsible for these effects is not known, and indeed several factors, including enzyme inhibition and tissue damage, may be involved [68]. The CNS effects mainly impact children under 5 years of age, since lead does not pass the blood-brain barrier as easily in adults. Asymptomatic exposure to low doses of lead in young children (toddlers) can lead to permanent decreases in IQ.
- **Kidney Damage:** Kidney damage characterized by atrophy of the renal tubules is a well-established effect of lead poisoning. The damage is associated with elevated levels of amino acids, sugar, and phosphate in the urine [68]. Gout caused by kidney damage has also been associated with lead poisoning, the consumption of illicitly distilled whiskey (moonshine) being a common cause. Concentrations of lead in moonshine often exceed 10 ppm due to the use of lead in the distillation apparatus [73]. Waldron and Stöfen [68] cite one example of

a hospital in which 37 of 43 cases of gout admitted in 1967 were caused by the consumption of lead-contaminated liquor.

- **Effects on Children and Reproduction:** From the standpoint of human health, the greatest danger of lead poisoning is undoubtedly to young children, particularly those living in urban areas. Neurological damage caused by the poisoning of such children may be permanent and can result in impaired physical as well as mental development. Even the unborn fetus is not protected from the effects of lead poisoning. There is evidence that the brain of the fetus is much more sensitive to lead poisoning than the brain of the infant or young child, and lead has repeatedly been shown to cause birth defects in experimental animals [73]. Although there is little information to indicate that lead has a teratogenic effect on humans, exposure of pregnant women to lead can induce miscarriages and stillbirths [68].

Most of the body burden of lead is found in the bones, and although the lead content of bones can now be assayed using noninvasive x-ray fluorescence techniques [74], historically it has been much more practical to relate health effects on living persons to the level of lead in the blood, PbB. Table 1 summarizes the lowest PbB levels associated with various health effects as summarized by the EPA [73] and CDC [75].

Toxic Chemical Risks. Table 1 Lowest PbB levels associated with health effects in humans

Lead in blood ($\mu\text{g L}^{-1}$)	Effect	Population group
100	ALA-d inhibition	Children and adults
	Adverse effects on intelligence, hearing, and growth	Children
100–150	Interference with vitamin D metabolism	Children
150–200	Protoporphyrin elevation in red blood cells	Women and children
250–300	Protoporphyrin elevation in red blood cells	Adult males
400	Increased urinary ALA excretion	Children and adults
400	Anemia	Children
400	Coproporphyrin elevation	Adults and children
500	Anemia	Adults
500–600	Cognitive deficits	Children
500–600	Peripheral neuropathies	Adults and children
800–1,000	Encephalopathic symptoms	Children
1,000–1,200	Encephalopathic symptoms	Adults

PbB levels as low as $100 \mu\text{g L}^{-1}$ do not cause distinctive symptoms, but are associated with decreased intelligence and impaired neurobehavioral development, decreased stature or growth, decreased hearing acuity, and decreased ability to maintain a steady posture. The activity of δ -aminolaevulinic acid dehydratase (ALA-d), an enzyme involved in heme metabolism, is also adversely affected at a PbB level of $100 \mu\text{g L}^{-1}$. Interference with vitamin D metabolism occurs at PbB levels of $100\text{--}150 \mu\text{g L}^{-1}$, and increased concentrations of protoporphyrin, a substrate needed for heme synthesis, and ALA begin to appear in red blood cells and urine, respectively, at slightly higher PbB levels. Actual decreases in hemoglobin, somewhat loosely referred to in Table 1 as anemia, occur at PbB

levels of about $400\text{--}500 \mu\text{g L}^{-1}$. Based on this evidence, the Centers for Disease Control have concluded that children's PbB levels should not exceed $100 \mu\text{g L}^{-1}$ [75].

Unfortunately a great many children as well as adults in the United States have PbB levels greater than $100 \mu\text{g L}^{-1}$. For example, a study of urban children in seven different US cities between 1967 and 1970 revealed that about 29% had blood levels over $400 \mu\text{g L}^{-1}$, and almost 9% had blood lead levels over $500 \mu\text{g L}^{-1}$ [68]. A blood lead level of $500\text{--}800 \mu\text{g L}^{-1}$ is considered to be the threshold level of classical lead poisoning [76]. At the present time in the United States $\sim 250,000$ children less than 6 years old have blood lead levels exceeding $100 \mu\text{g L}^{-1}$ [77].

Now that leaded gasoline has been phased out, the principal sources of exposure to lead are deteriorating lead-based paint, lead-contaminated dust, and lead-contaminated residential soil [78]. Of these three, the historical use of lead-based paint has proven to be the most problematic. Lead was used for many years in paints as a pigment, and in oil paints lead naphthenate is used as a drying agent. In 1918, 40% of all painters were estimated to have symptoms of lead poisoning because of their contact with such paint [79]. The most serious problem associated with the use of lead in paints, however, has been the poisoning of small children who eat paint chips or teethe on windowsills or other surfaces that have been painted with lead-based paints. Dust from deteriorating paint may also contaminate windowsills, carpets, furniture, and toys and be ingested by children who touch the dust and subsequently put their fingers in their mouths. Between 1954 and 1967, a total of 2,038 children were treated for lead poisoning in New York City alone, and of these 128 died [79]. On January 1, 1973, the Food and Drug Administration (FDA) banned the use of paint containing over 0.5% lead on residential surfaces accessible to children [68], and effective February 27, 1978, paint containing over 0.06% lead was banned as hazardous for use in residences, schools, hospitals, parks, playgrounds, and public buildings under the aegis of the Consumer Product Safety Act. The only exceptions specifically written into the 1977 legislation were paints and coatings for motor vehicles and boats. The problem of poisoning from lead-based paints remains a serious public health problem, however, because many older buildings contain lead-based paints. A 1988 Public Health Service report revealed that 52% of American homes

had layers of lead-based paint on their walls and woodwork [74]. Today lead poisoning is the number one environmental health risk for children in the United States, and 80% of lead poisonings are caused by lead-based paint in homes and apartments built before 1978 [80].

Naively one might assume that simply painting over lead-based paint would solve the problem, but according to the EPA, that is “not enough” [78].

- To permanently remove lead hazards, you must hire a certified lead “abatement” contractor. Abatement (or permanent hazard elimination) methods include removing, sealing, or enclosing lead-based paint with special materials.... Always hire a person with special training for correcting lead problems – someone who knows how to do this work safely and has the proper equipment to clean up thoroughly [78].

Short of having the lead-based paint removed by a certified contractor, people who find themselves living in housing units painted with lead-based paint are advised to take the following steps [78]:

- If you rent, notify your landlord of peeling or chipping paint.
- Clean up paint chips immediately.
- Clean floors, window frames, windowsills, and other surfaces weekly. Use a mop, sponge, or paper towel with warm water and a general all-purpose cleaner or a cleaner made specifically for lead. Thoroughly rinse sponges and mop heads after cleaning dirty or dusty areas.
- Wash children’s hands often, especially before they eat and before nap time and bedtime.
- Keep play areas clean. Wash bottles, pacifiers, toys, and stuffed animals regularly.
- Keep children from chewing windowsills or other painted surfaces.
- Clean or remove shoes before entering your home to avoid tracking in lead from soil.

Other Chemicals

Almost 90% of all poison exposures occur in the home, and after lead there is a rather long list of toxic chemicals to which people have been exposed, in most (~87%) cases unintentionally. Seventy-five

percent of poison exposures involve ingestion of a toxic substance, and children less than 6 years old are the victims in roughly half the incidents, although they account for only 2% of the poison fatalities. In almost 40% of the cases involving children, cosmetics and personal care products, cleaning compounds, analgesics, and plants (e.g., holly berries and mistletoe berries) are the source of the toxin(s). Although adults account for a small percentage of poison exposures, they account for more than 75% of all fatalities related to poison exposure [81]. The following examples illustrate several facets of the poison exposure problem.

1. Antibacterials found in many cosmetic and personal care products: Triclosan is a chlorinated organic bactericide used in soaps, deodorants, toothpastes, shaving creams, mouthwashes, and cleaning supplies and is infused in consumer products such as kitchen utensils, toys, bedding, socks, and trash bags [82]. Its use for 40 years has been regulated by three different federal agencies, the FDA, EPA, and Consumer Product Safety Commission. Remarkably, the FDA has been working on the development of rules for the use of triclosan for much of that time but has yet to promulgate any [83]. Japan and Canada, on the other hand, have banned triclosan in consumer products.

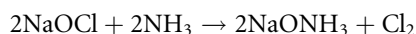
What is the problem? Concerns about the effects of triclosan on human health are related to several issues. First, several studies have linked triclosan to allergic reactions in children [84–87]. Second, triclosan reacts with chlorine in water to produce chloroform, which the EPA classifies as a probable human carcinogen, and smaller amounts of other chlorinated compounds that, in the presence of sunlight, are converted to dioxin [88–91], some forms of which are highly toxic and very potent endocrine disruptors. So widespread is the use of triclosan that it has been found in the urine of 75% of people tested by the CDC [83] and even in human breast milk [92, 93].

Given this information, one might wonder why the US government has not taken some action to restrict the use of triclosan, but there is another side to the story. In commenting on the presence of triclosan in breast milk, Dayan [92], for example,

concludes, “There is no evidence to indicate that the presence of a miniscule amount of triclosan in breast milk presents a risk to babies,” and while it is true that triclosan reacts with chlorine in tap water to produce chloroform, the concentrations of chloroform reported in the study by Rule et al. [94] were smaller than the amounts often present in chlorinated drinking water [82]. With respect to allergic reactions, Campbell and Zirwas [86] acknowledge that, “Several cases of contact dermatitis secondary to triclosan have been reported” but that, “Allergy to triclosan-based products is uncommon,” and Schena et al. [85], in a study of patients with chronic eczema, concluded, “Triclosan is well tolerated and has a very low sensitizing potential even in high-risk patients affected by eczema.”

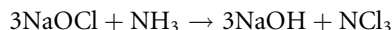
There is no question that very large numbers of people (arguably almost everyone) in the United States are exposed to triclosan, but is there a health risk associated with exposure to this chemical? It seems that the jury is still out on this question. And would the health risk of bacterial infection be heightened if use of triclosan were banned? Remarkably, a study carried out by the University of Michigan concluded, “Washing hands with an antibacterial soap was no more effective in preventing infectious illness than plain soap. Moreover, antibacterial soaps at formulations sold to the public do not remove any more bacteria from the hands during washing than plain soaps” [95].

2. Dangerous mixtures of household chemicals: Without realizing the danger, people sometimes mix bleach with ammonia. The resulting chemical reactions can produce a real witches’ brew of toxic gases. Bleach is just a dilute (typically 6%) aqueous solution of sodium hypochlorite (NaOCl). Ammonia (NH₃) is a gas, and when dissolved in water exists as a mixture of the dissolved gas and ammonium ion (NH₄⁺). When roughly equal amounts of bleach and ammonia are combined the following reaction occurs:

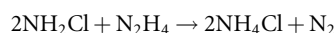
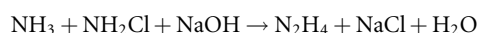
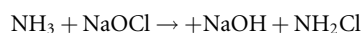


Chlorine gas, Cl₂, is a highly corrosive gas that can do serious damage to the lungs, nasal passages, and trachea when inhaled. It was used as a chemical weapon during World War I and by Nazi Germany in World War II. When more

bleach is added than ammonia the following reaction occurs:

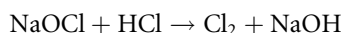


NaOH is a very strong base and caustic to soft tissue. NCl₃, nitrogen trichloride, is highly toxic and sufficiently volatile that it is likely to explode in the face of the unsuspecting chemical mixer. And finally, if ammonia is added in excess of the amount of bleach, the following series of reactions may occur:

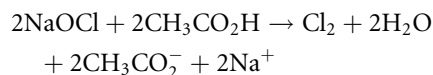


The third reaction is very exothermic (releases a large amount of heat) and typically leads to an explosion.

Another common mistake is to mix bleach with either toilet bowl cleaners or vinegar. People presumably do this because they think mixing the two will produce a better cleaner and disinfectant. In both cases, one of the products is Cl₂, which, as already noted, is a very toxic gas. Mixing bleach and toilet bowl cleaner, which contains hydrochloric acid (HCl), leads to the following reaction:



Mixing bleach with vinegar, which contains acetic acid (CH₃CO₂H), leads to the following reaction:



In either case, the result could be very painful and possibly fatal for the unwary mixer of chemicals.

Future Directions

Given the fact that most exposures to toxic chemicals are associated with lifestyle issues, it is provocative to ask what lies ahead. Clearly the prevalence of smoking has been on the decline in the United States, and if the trend apparent in Fig. 1 is extrapolated another 50 years, only a small percentage of adults will be smoking by that time. The risks associated with smoking are well established. The challenges to reducing the societal

costs associated with smoking now lie in the realms of education and social science. How far the percentage of smokers can be reduced through education and public awareness efforts is unclear at this time.

The risks associated with alcohol consumption are also well defined, and because alcohol consumption is much more socially acceptable than smoking, it is unlikely that the costs associated with drinking will be substantially reduced in the years ahead.

Abuse of drugs is a major public health problem whose solution, if there is one, will lie in the realm of education and social science. Despite the best efforts of law enforcement agencies, it is apparent that if people want illegal drugs, they will get them, and if they want prescription drugs without a prescription, that can happen too. The solution will require reducing the desire/need for the artificial euphoria associated with use of drugs such as opioids, cocaine, and heroin. Drug abuse is a difficult and complex problem, and particularly in the case of opioids (Fig 3) with no apparent resolution in sight.

Obesity is a serious problem in the United States, and awareness of the magnitude of the problem has been slow to emerge. Trends in recent years (Fig. 7) are not encouraging. Education and public awareness of the costs of obesity will help, but much of the underlying problem is the simple fact that people like to eat.

Lead has been historically and remains today a major environmental pollutant. The phasing out of leaded gasoline has been a very positive step, but now there is the legacy of the many homes and apartments that were painted with lead-based paint prior to 1978. Eventually these will be taken out of service, but the process has been slow because of the cost and hazards associated with renovating interiors painted with lead-based paint.

Numerous unintentional exposures to toxic chemicals occur in and around the home. Improved education and public awareness of the hazards of certain chemicals can help reduce the frequency of these exposures. The presence of potential toxins in consumer products is an issue that must be addressed by government agencies such as the FDA, EPA, and Consumer Product Safety Commission. In some cases a judgment must be made between the benefits and the risks. In the United States, 35,000–40,000 people die in automobile accidents each year, but there is no movement to take cars off the roads. While the benefits

and costs associated with automobile use are arguably well defined, such is not the case with triclosan, for which both the benefits and risks are in dispute.

Bibliography

Primary Literature

1. WHO (1987) Principles for the safety assessment of food additives and contaminants in food. WHO Press, Geneva
2. CDC (2009) Glossary of terms. <http://www.atsdr.cdc.gov/glossary.html>. Accessed 16 Mar, 2011
3. Hoffmann D, Hoffmann I (2001) The changing cigarette: chemical studies and bioassays. In: Burns DM, Benowitz NL (eds) Risks associated with smoking cigarettes with low machine-measured yields of tar and nicotine. U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health, National Cancer Institute, Washington, DC, pp 159–191
4. Hoffmann D, Hoffmann I (1998) Tobacco smoke components. *Beitr Tabakforsch Int* 18:49–52
5. Hoffmann D, Hoffmann I, El-Bayoumy K (2001) The less harmful cigarette: a controversial issue. A tribute to Ernst L. Wynder. *Chem Res Toxicol* 14(7):767–790
6. Baker RR, da Silva JRP, Smith G (2004) The effect of tobacco ingredients on smoke chemistry. Part I: Flavourings and additives. *Food Chem Toxicol* 42:S3–S37
7. Staff (1990) Proposed identification of inorganic arsenic as a toxic air contaminant. California Air Resources Board and Department of Health Services, Sacramento
8. Korte JE, Hertz-Picciotto I, Schulz MR, Ball LM, Duell EJ (2000) The contribution of benzene to smoking-induced leukemia. *Environ Health Perspect* 108(4):333–339
9. Staff (2004) The health consequences of smoking: a report of the surgeon general. Centers for Disease Control and Prevention, Atlanta, GA
10. Denissenko MF, Pao A, Tang MS, Pfeifer GP (1996) Preferential formation of benzo a pyrene adducts at lung cancer mutational hotspots in P53. *Science* 274(5286):430–432
11. Kaiserman MJ, Rickert WS (1992) Carcinogens in tobacco-smoke – benzo a pyrene from Canadian cigarettes and cigarette tobacco. *Am J Public Health* 82(7):1023–1026
12. Greenblatt MS, Bennett WP, Hollstein M, Harris CC (1994) Mutations in the P53 tumor-suppressor gene – clues to cancer etiology and molecular pathogenesis. *Cancer Res* 54(18):4855–4878
13. Lane DP (1992) Cancer – P53, guardian of the genome. *Nature* 358(6381):15–16
14. EPA (1980) Ambient water quality for cadmium. EPA 440/5-80-025, Washington, DC
15. Kasuya M et al (1992) Water pollution by cadmium and the onset of Itai-Itai disease. *Water Sci Technol* 25(11):149–156
16. Olsson IM et al (2002) Cadmium in blood and urine – Impact of sex, age, dietary intake, iron status, and former smoking – Association of renal effects. *Environ Health Perspect* 110(12): 1185–1190

17. Paschal DC et al (2000) Exposure of the US population aged 6 years and older to cadmium: 1988–1994. *Arch Environ Contam Toxicol* 38(3):377–383
18. Bokhoven C, Niessen HJ (1961) Amounts of oxides of nitrogen and carbon monoxide in cigarette smoke, with and without inhalation. *Nature* 192(480):458
19. Kaufman DW, Helmrich SP, Rosenberg L, Miettinen OS, Shapiro S (1983) Nicotine and carbon-monoxide content of cigarette-smoke and the risk of myocardial-infarction in young men. *N Engl J Med* 308(8):409–413
20. Counts ME, Hsu FS, Laffoon SW, Dwyer RW, Cox RH (2004) Mainstream smoke constituent yields and predicting relationships from a worldwide market sample of cigarette brands: ISO smoking conditions. *Regul Toxicol Pharmacol* 39(2):111–134
21. Bateman DN (2003) Carbon monoxide. *Medicine* 31(10):41–42
22. Townsend CL, Maynard RL (2002) Effects on health of prolonged exposure to low concentrations of carbon monoxide. *Occup Environ Med* 59(10):708–711
23. Baker RR (2006) The generation of formaldehyde in cigarettes – overview and recent experiments. *Food Chem Toxicol* 44(11):1799–1822
24. Baker RR, Coburn S, Liu C (2006) The pyrolytic formation of formaldehyde from sugars and tobacco. *J Anal Appl Pyrol* 77(1):12–21
25. Li S, Banyasz JL, Parrish ME, Lyons-Hart J, Shafer KH (2002) Formaldehyde in the gas phase of mainstream cigarette smoke. *J Anal Appl Pyrol* 65(2):137–145
26. Albert RE et al (1982) Gaseous formaldehyde and hydrogen-chloride induction of nasal cancer in the rat. *J Natl Cancer Inst* 68(4):597–603
27. Blair A, Saracci R, Stewart PA, Hayes RB, Shy C (1990) Epidemiologic evidence on the relationship between formaldehyde exposure and cancer. *Scand J Work Environ Health* 16(6):381–393
28. Casanova M, Morgan KT, Gross EA, Moss OR, Heck HD (1994) DNA-protein cross-links and cell replication at specific sites in the nose of F344 rats exposed subchronically to formaldehyde. *Fundam Appl Toxicol* 23(4):525–536
29. Olsen JH et al (1984) Occupational formaldehyde exposure and increased nasal cancer risk in man. *Int J Cancer* 34(5):639–644
30. Horvath EP, Anderson H, Pierce WE, Hanrahan L, Wendlick JD (1988) Effects of formaldehyde on the mucous-membranes and lungs – a study of an industrial-population. *JAMA* 259(5):701–707
31. Liteplo RG, Meek ME (2003) Inhaled formaldehyde: exposure estimation, hazard characterization, and exposure-response analysis. *J Toxicol Environ Health B Crit Rev* 6(1):85–114
32. Rodgman A, Perfetti TA (2008) The chemical components of tobacco and tobacco smoke. CRC Press, Boca Raton, 1840 pp
33. Hathaway GJ, Proctor NH, Hughes JP, Fischman ML (1991) Proctor and Hughes' chemical hazards of the workplace. Van Nostrand Reinhold, New York, NY
34. Argentin G, Cicchetti R (2004) Genotoxic and antiapoptotic effect of nicotine on human gingival fibroblasts. *Toxicol Sci* 79(1):75–81
35. Campain JA (2004) Nicotine: potentially a multifunctional carcinogen? *Toxicol Sci* 79(1):1–3
36. Hoffmann D, Brunnemann KD, Prokopczyk B, Djordjevic MV (1994) Tobacco-specific N-nitrosamines and areca-derived N-nitrosamines – chemistry, biochemistry, carcinogenicity, and relevance to humans. *J Toxicol Environ Health* 41(1):1–52
37. Pfeifer GP et al (2002) Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. *Oncogene* 21(48):7435–7451
38. Hoffmann D, Hoffmann I (1997) The changing cigarette, 1950–1995. *J Toxicol Environ Health* 50(4):307–364
39. CDC (2011) Toll of tobacco in the United States. University of Pennsylvania Center for Interdisciplinary Research on Nicotine Addiction, Philadelphia, PA. http://www.med.upenn.edu/cirna/pdf/USA_Figures.pdf. Accessed 26 Apr 2011
40. CDC (2011) Annual smoking-attributable mortality, years of potential life lost, and economic costs – United States, 1995–1999. http://www.cdc.gov/tobacco/data_statistics/mmwr/byyear/2002/mm5114a2/highlights.htm. Accessed 6 Mar 2011
41. CDC (2011) Economic facts about U.S. tobacco production and use. http://www.cdc.gov/tobacco/data_statistics/fact_sheets/economics/econ_facts/#costs. Accessed 7 Mar 2011
42. CDC (2011) Trends in current cigarette smoking among high school students and adults, United States, 1965–2009. Centers for Disease Control and Prevention, Atlanta, GA. http://www.cdc.gov/tobacco/data_statistics/tables/trends/cig_smoking/index.htm. Accessed 26 Apr 2011
43. Courtwright DT (2001) Forces of habit: drugs and the making of the modern world. Harvard University Press, Cambridge, MA
44. Lord JL et al (1987) Coming to grips with alcoholism. *US News World Rep* 103(22):57–62
45. Anonymous (2011) Alcohol abuse statistics. http://www.about-alcohol-abuse.com/Alcohol_Abuse_Statistics.html. Accessed 8 Mar 2011
46. Harford TC (1992) Family history of alcoholism in the United States – prevalence and demographic characteristics. *Br J Addict* 87(6):931–935
47. Anonymous (2010) The party's over. *Nature* 468(7323):475
48. Anonymous (2011) Alcoholism. http://www.addiction-help-line.com/alcohol_addiction.html. Accessed 9 Mar 2011
49. Anthenelli RM, Schuckit MA (1993) Affective and anxiety disorders and alcohol and drug dependence: diagnosis and treatment. *J Addict Disord* 12:73–87
50. Goodwin DW (1985) Alcoholism and genetics – the sins of the fathers. *Arch Gen Psychiatry* 42(2):171–174
51. Paulozzi LJ, Ryan GW (2006) Opioid analgesics and rates of fatal drug poisoning in the United States. *Am J Prev Med* 31(6):506–511
52. CDC (2011) CDC WONDER. <http://wonder.cdc.gov/>. Accessed 11 Mar 2011

53. Paulozzi LJ (2008) Trends in unintentional drug overdose deaths. <http://www.hhs.gov/asl/testify/2008/03/t20080312b.html>. Accessed 11 Mar 2011
54. Warner ME, Chen LH, Makuc DM (2009) Increase in fatal poisonings involving opioid analgesics in the United States, 1999–2006. Centers for Disease Control and Prevention, Atlanta, GA
55. CDC (2011) Unintentional drug poisoning in the United States. <http://www.cdc.gov/HomeandRecreationalSafety/pdf/poison-issue-brief.pdf>. Accessed 11 Mar 2011
56. Anonymous (2011) Opioids. <http://www.drug-addiction.com/opioids.htm>. Accessed 11 Mar 2011
57. NIDA (2010) Cocaine: abuse and addiction. U.S. Department of Health and Human Services, National Institutes of Health, Research Triangle Park, NC
58. Flynn PM, Joe GW, Broome KM, Simpson DD, Brown BS (2003) Looking back on cocaine dependence: reasons for recovery. *Am J Addict* 12(5):398–411
59. Johnston LD, O'Malley PM, Bachman JG, Schulenberg JE (2010) Marijuana use is rising; ecstasy use is beginning to rise; and alcohol use is declining among U.S. teens. <http://www.monitoringthefuture.org>. Accessed 11 Mar 2011
60. NIDA (2005) Heroin abuse and addiction. U.S. Department of Health and Human Services, National Institutes of Health, Research Triangle Park, NC
61. Anonymous (2011) Morphine. <http://en.wikipedia.org/wiki/Morphine>. Accessed 12 Mar 2011
62. Anonymous (2011) Heroin. <http://en.wikipedia.org/wiki/Heroin>. Accessed 12 Mar 2011
63. CDC (2002) Drug-associated HIV transmission continues in the United States. Centers for Disease Control and Prevention, Atlanta, GA
64. Anonymous (2004) Results from the 2003 national survey on drug use and health: national findings. Office of Applied Studies, Research Triangle Park, NC
65. NCHS (2008) Prevalence of overweight, obesity and extreme obesity among adults: United States, trends 1976–80 through 2005–2006. National Center for Health Statistics, Hyattsville, MD
66. NCHS (2004) Mean body weight, height, and body mass index, United States 1960–2002. National Center for Health Statistics, Hyattsville, MD
67. NHLBI (1998) Clinical guidelines on the identification, evaluation, and treatment of overweight and obesity in adults. National Heart, Lung, and Blood Institute, Research Triangle Park, NC
68. Waldron HA, Stöfen D (1974) Sub-clinical lead poisoning. Academic, New York
69. Gilfillan SC (1965) Lead poisoning and the fall of Rome. *J Occup Environ Med* 7(2):53–60
70. IPIECA (2010) Leaded gasoline phase-out. <http://www.ipieca.org/topic/operations-and-fuels/leaded-gasoline-phase-out>. Accessed 12 Mar 2011
71. McDonald J (2007) China's lead problems go beyond toys. http://www.usatoday.com/money/economy/2007-08-15-3066711946_x.htm. Accessed 12 Mar 2011
72. Annet JL (1983) Trends in the blood lead levels of the U.S. population. In: Rutter M, Jones RR (eds) *Lead versus health*. John Wiley and Sons, Chichester and New York, pp 33–58
73. EPA (1980) Ambient water quality for lead. EPA-440/4-80-057. Environmental Protection Agency, Washington, DC
74. Blackman A (1991) Controlling a childhood menace. *Time* 137(8):68–69
75. CDC (1991) Preventing lead poisoning in young children. Centers for Disease Control and Prevention, Atlanta, GA
76. Patterson CC (1965) Contaminated and natural lead environments and man. *Arch Environ Health* 11:344–360
77. CDC (2011) Lead. <http://www.cdc.gov/nceh/lead/>. Accessed 13 Mar 2011
78. EPA (2011) Lead in paint, dust, and soil. <http://www.epa.gov/lead/>. Accessed 12 Mar 2011
79. Craig PP, Berlin E (1971) The air of poverty. *Environment* 13(5):2–9
80. Anonymous (2011) Frequently asked questions. <http://www.leadpro.com/faq.html>. Accessed 12 Mar 2011
81. Litovitz TL et al (2001) 2000 annual report of the American association of poison control centers toxic exposure surveillance system. *Am J Emerg Med* 19(5):337–395
82. Anonymous (2011) Triclosan. <http://en.wikipedia.org/wiki/Triclosan>. Accessed 13 Mar 2011
83. Layton L (2010) FDA says studies on triclosan, used in sanitizers and soaps, raise concerns. <http://www.washingtonpost.com/wp-dyn/content/article/2010/04/07/AR2010040704621.html>. Accessed 13 Mar 2011
84. Wong CSM, Beck MH (2001) Allergic contact dermatitis from triclosan in antibacterial handwashes. *Contact Dermat* 45(5):307–307
85. Schena D, Papagrigoraki A, Girolomoni G (2008) Sensitizing potential of triclosan and triclosan-based skin care products in patients with chronic eczema. *Dermatol Ther* 21: S35–S38
86. Campbell L, Zirwas MJ (2006) Triclosan. *Dermatitis* 17(4): 204–207
87. Bhutani T, Jacob SE (2009) Triclosan: a potential allergen in suture-line allergic contact dermatitis. *Dermatol Surg* 35(5): 888–889
88. Mezcuca M et al (2004) Evidence of 2,7/2,8-dibenzodichloro-p-dioxin as a photodegradation product of triclosan in water and wastewater samples. *Anal Chim Acta* 524(1–2):241–247
89. Latch DE et al (2005) Aqueous photochemistry of triclosan: Formation of 2,4-dichlorophenol, 2,8-dichlorodibenzo-p-dioxin, and oligomerization products. *Environ Toxicol Chem* 24(3):517–525
90. Latch DE, Packer JL, Arnold WA, McNeill K (2003) Photochemical conversion of triclosan to 2,8-dichlorodibenzo-p-dioxin in aqueous solution. *J Photochem Photobiol A Chem* 158(1): 63–66
91. Aranami K, Readman JW (2007) Photolytic degradation of triclosan in freshwater and seawater. *Chemosphere* 66(6): 1052–1056

92. Dayan AD (2007) Risk assessment of triclosan Irgasan (R) in human breast milk. *Food Chem Toxicol* 45(1): 125–129
93. Allmyr M, Adolfsson-Erici M, McLachlan MS, Sandborgh-Englund G (2006) Triclosan in plasma and milk from Swedish nursing mothers and their exposure via personal care products. *Sci Total Environ* 372(1):87–93
94. Rule KL, Ebbett VR, Vikesland PJ (2005) Formation of chloroform and chlorinated organics by free-chlorine-mediated oxidation of triclosan. *Environ Sci Technol* 39(9):3176–3185
95. Anonymous (2007) Plain soap as effective as antibacterial but without the risk. <http://www.physorg.com/news106418144.html>. Accessed 14 Mar 2011

Books and Reviews

- Higgins TE, Sachdev JA, Engleman SA (2010) Toxic chemicals: risk prevention through use reduction. CRC Press, Boca Raton
- Nielsen E, Ostergaard G, Larsen JC (2008) Toxicological risk assessment of chemicals: a practical guide. Informa Healthcare, London
- Penningroth S (2010) Essentials of toxic chemical risk: science and society. CRC Press, Boca Raton
- Richardson ML (ed) (1996) Risk reduction: chemicals and energy into the 21st century. Taylor and Francis, London

Toxic Organic Chemicals

SIMANGA GAMA¹, JON A. ARNOT², DON MACKAY¹

¹Canadian Centre for Environmental Modelling and Chemistry, Trent University, Peterborough, ON, Canada

²Department of Physical and Environmental Sciences, University of Toronto Scarborough, Toronto, ON, Canada

Article Outline

Glossary
 Definition of the Subject
 Introduction
 Properties and Characteristics of Toxic Organic Chemicals
 The Stockholm Convention on Persistent Organic Pollutants (POPs)
 Chemical Classes

Conclusions
 Future Directions
 Bibliography

Glossary

Bioaccumulation The phenomenon similar to bioconcentration but including uptake from food as well as uptake from the ambient environment. This is expressed as a bioaccumulation factor (BAF) and generally applies to organisms in the environment. The BAF is the steady-state ratio of the chemical concentration in the organism to that in the environment.

Bioconcentration The phenomenon by which an organism, such as a fish, absorbs chemical from its ambient environment of water or air by respiratory uptake and/or dermal absorption. The steady-state ratio of the chemical concentration in the organism to the chemical concentration in its ambient environment of water or air is the bioconcentration factor, (BCF) and the BCF is usually measured in a laboratory test.

Biomagnification The ratio of the chemical concentration in the predator to that of the prey.

Hazard The inherent toxic potency of a chemical, usually expressed as the quantity or concentration of the substance necessary to elicit a defined adverse effect in the organism.

Mass Balance Model A mathematical description of the fate and transport of a chemical in the environment, usually in the form of a computer program. The model provides a complete accounting of all processes experienced by the chemical and is used to estimate environmental concentrations, persistence, and exposures. Models may also be used to forecast future changes in concentrations as a result of actions to reduce contamination.

Persistence The average time that a discharged chemical survives in the environment before it is degraded into another substance or substances. It may be expressed as a “half-life” by analogy to radioactive substances or as a residence time.

Risk The likelihood that there will be an adverse effect as a result of exposure to the chemical. Risk thus depends both on toxic potency (hazard) and the prevailing exposure.

Toxicity The phenomenon by which a chemical substance elicits an adverse effect on an exposed organism. The effect may be death (lethality) or a less severe effect such as a failure to reproduce, an increased vulnerability to predation or significant behavioral changes.

Definition of the Subject

Organic chemicals play an invaluable role in the modern lifestyle. They include pharmaceuticals, pesticides, plastics, fuels, solvents, explosives, surface coatings, adhesives, disinfectants, and fire retardants. From the perspective of conservation and sustainability, the preferred strategy is to use chemicals such that they perform their desired function, cause no unintended adverse effects, and hence leave no legacy of contamination. This is a fundamental component in the move toward “green chemistry” which also strives to reduce resource depletion, energy use, ozone depletion, and interference with natural biogeochemical cycles. Of the over 50 million chemicals that have been characterized, most are organic compounds. Some 100,000 are commercially produced in quantities large enough to raise concerns that they may become present in the environment in sufficient quantities and at sufficient concentrations to cause risks to the well-being of humans or other organisms. As a response to this concern, most regulatory agencies have listed chemicals of national concern, a typical number being in the hundreds to thousands. These “toxic organic chemicals” merit regulatory scrutiny and possible controls over synthesis and use. At the international level, the Stockholm Convention has listed over 20 substances as worthy of regulation or even bans. Most of these are synthetic chlorinated organic substances, but some naturally occurring and inadvertently produced substances are also of concern. In this entry, the chemical attributes or properties of these substances are described and criteria that dictate the level of concern about hazard and risk to environmental and human health are outlined. Selected classes of organic chemicals are then discussed in more detail.

Introduction

A consensus has emerged among the scientific and regulatory communities that it is a combination of properties

of chemical substances that dictates their designation as “toxic organic substances” and thus the need for their regulation. The result has been the identification of classes of chemicals as persistent organic pollutants (POPs) and persistent bioaccumulative and toxics (PBTs) and those that undergo long-range transport (LRT) on a global scale. Although numerical criteria have been developed for certain key chemical properties such as toxicity and persistence, it is apparent that there is a continuum of properties, thus there is no clear demarcation between for example, PBTs and non-PBTs. There are varying degrees of PBT-like character among “toxic organic chemicals.” It can be justifiably argued that all chemical substances are toxic if the administered dose or exposure is sufficient, that is, as stated over 500 years ago by Paracelsus, “it is the dose that makes the poison.” Accordingly, there is no clear demarcation between toxic and nontoxic substances: the demarcation must be on the basis of both potency as a toxic agent and the exposure experienced in the environment. The regulatory focus must be on ensuring the principles of sustainability and conservation by reducing exposure.

It is important to discriminate between the terms hazard and risk, since these terms are often wrongly used interchangeably. Hazard reflects the inherent properties of the chemical such as its toxic potency, regardless of the actual exposure. Risk reflects the probability of adverse effects as a function of both hazard and the actual exposure. A highly hazardous substance may pose a low risk and vice versa.

Properties and Characteristics of Toxic Organic Chemicals

Organic chemicals tend to fall into distinct classes based on their molecular structure, an example being the alcohols of increasing carbon chain length, methanol, ethanol, propanol, butanol, etc. There is often a systematic change in properties such as boiling point or vapor pressure in such homologous series. This feature is invaluable in that interpolation and modest extrapolation of properties is often possible. This is formalized in linear free energy relationships (LFER) and is exploited in the development of quantitative structure activity (or property) relationships (QSARs and QSPRs) that are widely used for property estimation purposes. As quantum chemical molecular

modeling computation methods become more reliable and accessible, it is clear that the key properties of chemical substances are amenable to a priori estimation from molecular structure.

A key LFER arises from a systematic increase in halogenation of a parent organic substance. For example, the series benzene, chlorobenzene, dichlorobenzene to hexachlorobenzene displays consistent changes in properties such as a decrease in vapor pressure and aqueous solubilities. This also occurs with bromine and to an extent with fluorine substitution and with increased methylation, for example, benzene, toluene, xylene, etc. These systematic property changes are exploited when later reviewing the properties of groups of chemicals such as the polychlorinated biphenyls (PCBs). The key properties as discussed by Mackay et al. [1] are addressed below.

Molecular structure and molar mass (g/mol) are obvious fundamental properties that reflect the constituent elements (molar mass) and their location in the chemical (molecular structure). It is necessary to discriminate between isomers – chemicals comprised of the same elements but with different structural or spatial patterns. For example, the four stereoisomers of hexachlorocyclohexane (HCH) are comprised of the exact same elements (six carbon, six chlorine, and six hydrogen atoms) contained in the same general structure (a cyclic ring of six carbon atoms, each with a bond to a chlorine and a hydrogen atom), but with different three-dimensional spatial patterns of these bonds. The seemingly minor differences in HCH isomer structure result in significant differences in toxicity.

Melting point and boiling point dictate the state of the pure chemical at atmospheric temperatures and pressure as either solids, liquids, or gases.

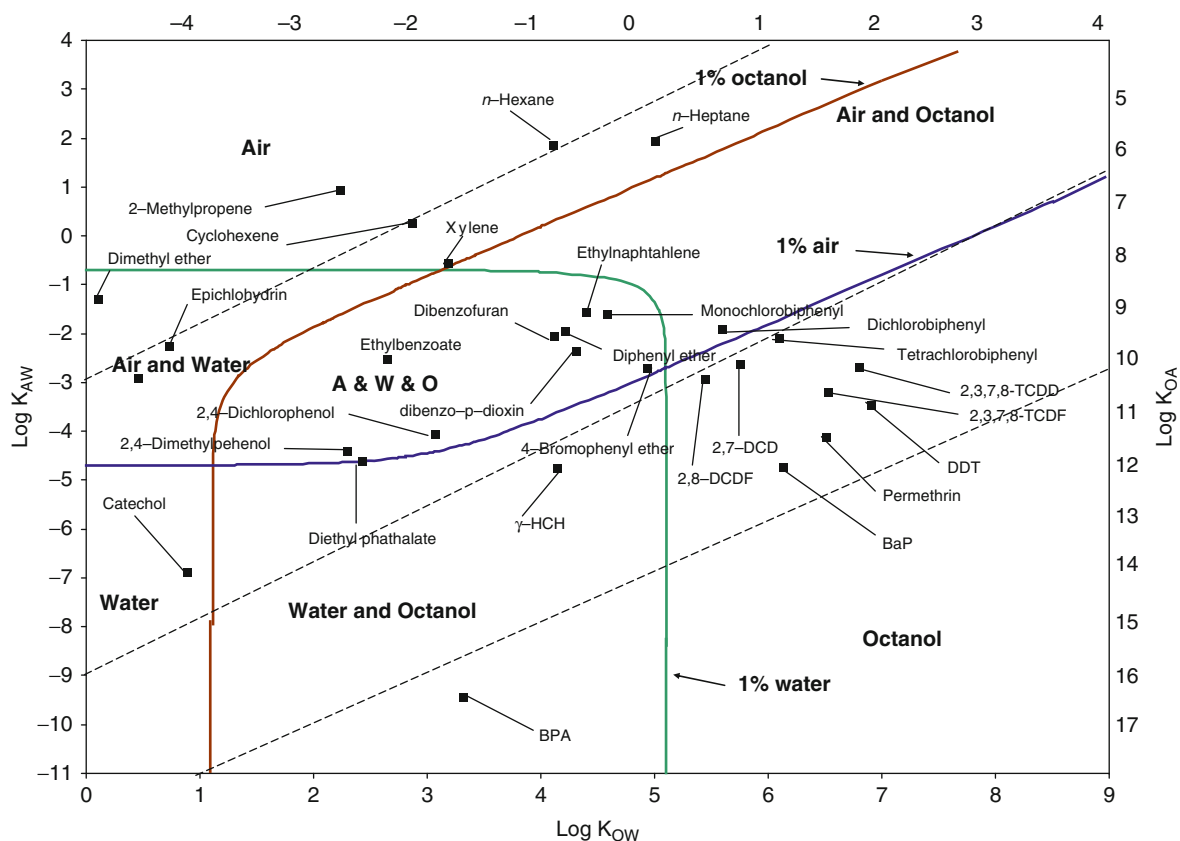
Chemical solubility and partitioning: the tendency of a substance to partition from its pure state into the atmosphere, water, and octanol, is expressed, respectively, as the saturation vapor pressure, aqueous solubility, and solubility in octanol. These properties are useful but are not always measurable. For example, the solubility of ethanol in water is not measurable because of miscibility. Octanol is widely accepted as a surrogate phase for partitioning into lipids (fats) in biota and for natural organic matter (OM) and organic carbon (OC) in soils and sediments. Since saturation conditions for organic chemicals rarely exist in the environment, it is

more convenient to use ratios of these three solubilities, that is, the three partition coefficients or ratios of concentrations, K_{AW} (air–water), K_{OW} (octanol–water), and K_{OA} (octanol–air). Clearly only two of these are independent since K_{OA} is K_{OW}/K_{AW} . The air–water partition coefficient K_{AW} is H/S^S_W or $(P^S/RT)/S^S_W$ where P^S is the saturation vapor pressure (Pa), R is the gas constant ($8.314 \text{ Pa m}^3/\text{mol K}$), T is absolute temperature (K), S^S_W is the solubility in water (mol/m^3), H is the Henry's law constant ($\text{Pa m}^3/\text{mol}$) and is P^S/S^S_W . For most environmental purposes, K_{OW} is assumed to be equivalent to the lipid–water partition coefficient. K_{OW} is also used to estimate partitioning to organic carbon from water (K_{OC}) as approximately $0.35 K_{OW}$ plus or minus a factor of three [2].

The K_{AW} , K_{OW} , K_{OA} properties thus largely dictate the equilibrium partitioning characteristics of the chemicals between air, water, soils, and sediments that contain organic carbon as well as biota in terms of their lipid content. A chemical space diagram provides a convenient depiction of these properties as plots of either $\log K_{AW}$ versus $\log K_{OW}$, (in which case the constant $\log K_{OA}$ values lie on a diagonal), or as $\log K_{AW}$ versus $\log K_{OA}$, with $\log K_{OW}$ values lying on a diagonal. Figure 1 is such a plot and shows the points corresponding to the series of chemicals representing particular chemical classes. For example, points for 1,1'-(2,2,2-trichloroethylidene)bis 4-chloro-benzene (DDT), benzo(a)pyrene, and PCBs are plotted showing the wide variation in partitioning properties which translate into differences in environmental fate. On this plot are lines corresponding to various estimated mass percentage partitioning between air, water, and octanol phases for an assumed volumetric proportion of these phases.

These simple relationships apply only if the molecule does not dissociate or ionize as applies to organic acids such as phenols at high pH and organic bases such as amines at low pH. For these substances, knowledge of the dissociation constant pK_A and the prevailing pH is essential.

Reactivity, expressed as a rate constant or half-life, is an essential property because it determines the persistence, that is, the residence time of the chemical in the environment. Persistence is important because to a first approximation the quantity of the chemical residing in the environment (kg) and hence the



Toxic Organic Chemicals. Figure 1

Chemical space plot of the air–water partition coefficient ($\log K_{AW}$) versus the octanol–water partition coefficient ($\log K_{OW}$) showing positions of chemical classes on the three media of air, water, and octanol. Octanol serves as a surrogate for organic matter in soils and sediments. Constant $\log K_{OA}$ values form the dashed diagonals on this plot of 1, 6, and 10 (from top to bottom) since $\log K_{OA}$ is $\log K_{OW} - \log K_{AW}$. The heavy lines represent 1% partitioning into each medium

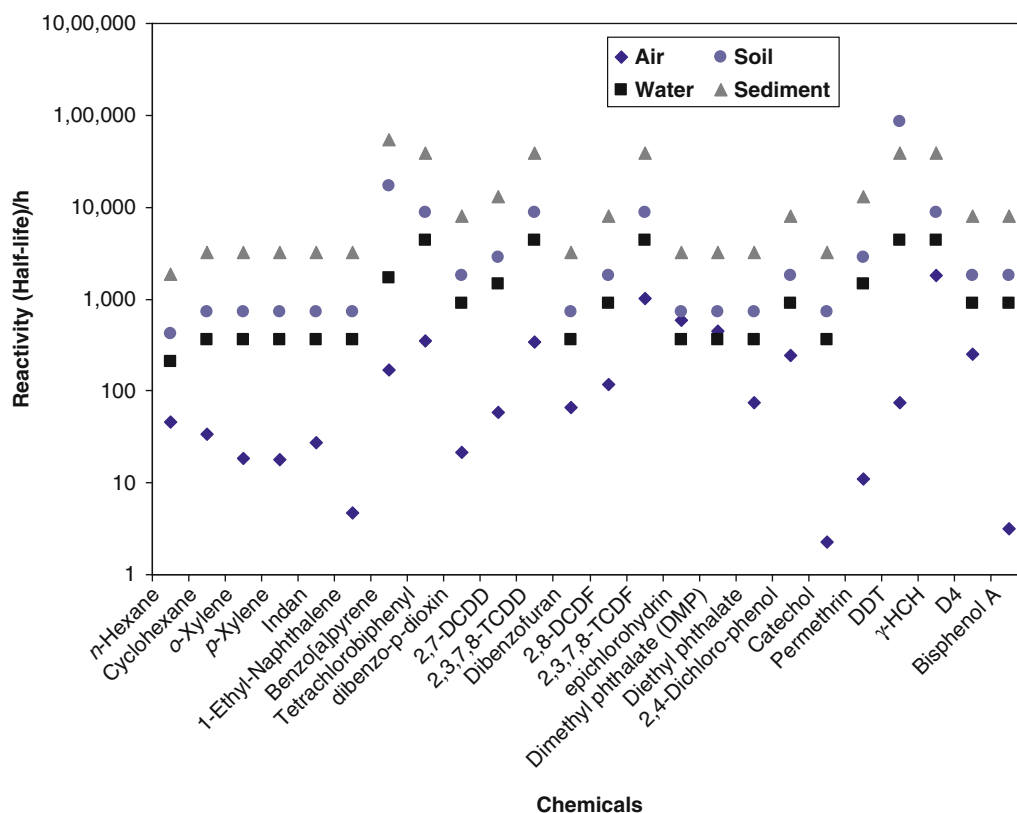
concentrations that exert toxicity (kg/m^3) are proportional to the discharge rate of chemical to the environment (kg/h) and to the chemical's residence time (h). Chemicals usually react following second-order kinetics

$$\text{Rate} = k_2 C_C C_R = k_1 C_C = 0.693 C_C / t_{1/2}$$

where k_2 is a second-order rate constant, C_C is chemical concentration, and C_R is the concentration of a reacting species such as hydroxyl radicals in the atmosphere [3]. It is often convenient to lump k_2 and C_R as a first-order rate constant k_1 (h^{-1}) but this assumes that C_R and k_2 are constant. In reality, both vary spatially and temporally, thus the half-life $t_{1/2}$ is also expected to vary. Despite this variability, persistence is such an important property that it is necessary

to provide estimates of half-lives of the chemical for all relevant media to which the chemical partitions. Some literature and databases exist on the half-lives of well-studied priority chemicals in air, water, soils, sediments, and in biota by metabolic conversions or biotransformation [1, 4–6]. There is, however, a paucity of data for other substances including “emerging” contaminants. The most common mechanisms are reaction with oxidizing species such as hydroxyl radicals, hydrolysis, photolysis (direct and indirect), biodegradation by microorganisms, and biotransformation in animals and plants.

Regulatory agencies have set half-life criteria for specific environmental media (for example, 60 days in water and 2 days in the atmosphere) [7]; however, in addition to medium-specific half-lives, the overall or



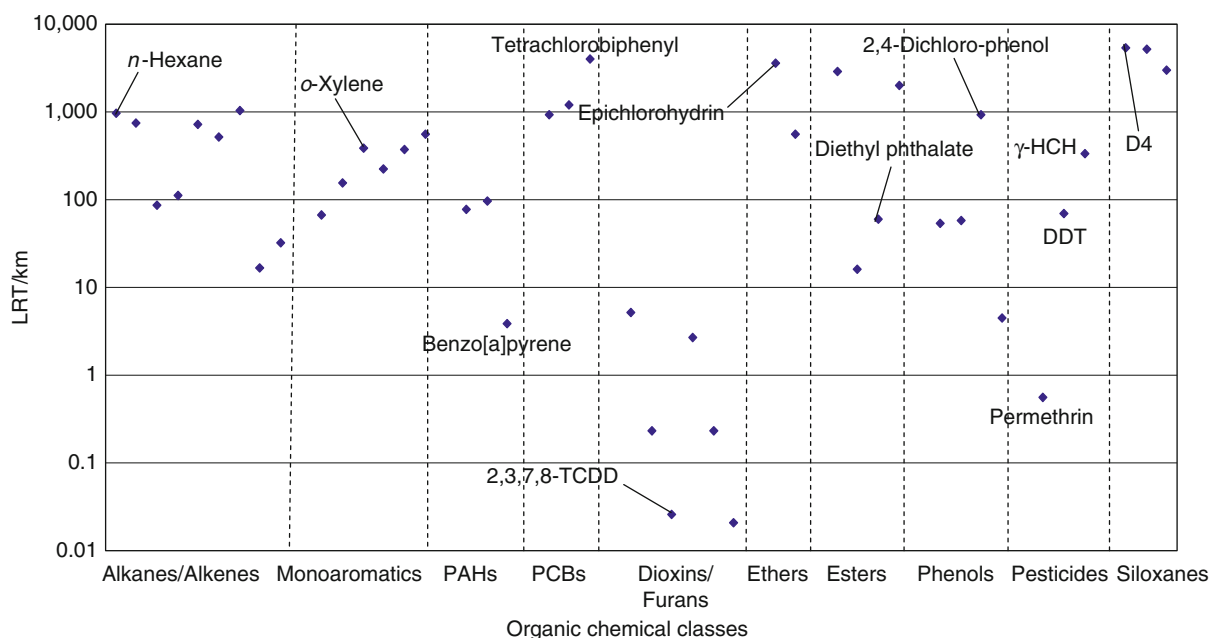
Toxic Organic Chemicals. Figure 2

Estimated half-lives (hours) of selected chemicals in air, water, soil, and sediment [6]

average persistence of a chemical in the environment is also important and is controlled by the relative quantities of chemical in each medium [8]. The relative quantities in each environmental medium, often referred to as the environmental fate and distribution, are largely a function of the partitioning properties of the chemical and the composition of the environment. In principle, the overall reaction rate constant (k_{OV}) is the weighted mean of the rate constants, the weighting being done according to the proportions in each medium. Computer programs such as the OECD Tool [9] are used to perform such calculations and yield an “overall persistence.” This quantity is essential in any assessment of sustainability because the year-to-year carryover of a chemical is controlled by overall persistence. In principle, an overall persistence ($P_{OV} = 0.693/k_{OV}$) of say 2 years implies that an appreciable fraction of the chemical discharged in year 1 will remain after 5 or more years. Indeed, this is the most important single property of an organic chemical,

hence their designation as persistent organic pollutants (POPs) [10]. See also the discussion by Gouin et al. [11], Mackay [12] and Scheringer [13]. Figure 2 illustrates the variation in reactivity of a selection of chemicals in each medium.

Potential for long-range transport (LRT) in air or oceans is increasingly recognized as a cause for concern because the adverse effects may be experienced at locations remote from the source [13]. Obvious regions of concern are the Arctic and Antarctic where local biota may be unacceptably contaminated and used as food by mammals including humans. Multimedia mass balance models have been developed that predict the influence of a chemical’s environmental phase distribution on its ability to be transported over long distances [14]. Persistence, and hence reactivity, of the chemical in air influences its potential to be transported over long distances in air. Multimedia models currently used to estimate LRT parameter include Globo-POP [15], RAIDAR [16], and the



Toxic Organic Chemicals. Figure 3

Long-range transport (LRT; km) in air estimated by Level II fate calculations using the RAIDAR 1.00 model [16].

The chemicals are categorized into general classes

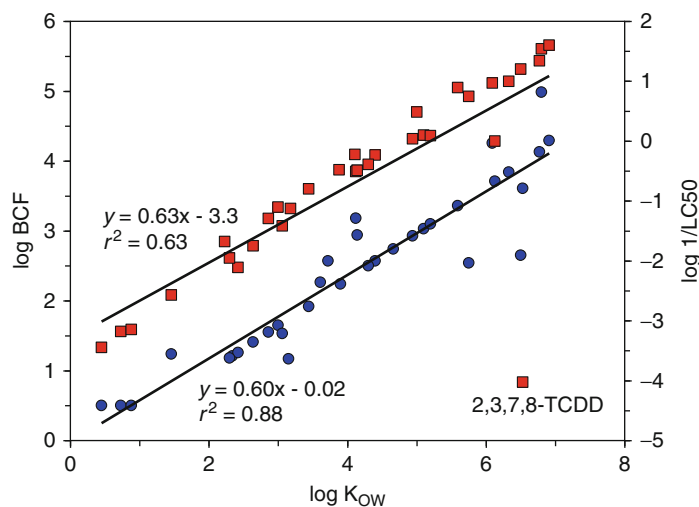
OECD Tool [9]. The models simulate LRT by incorporating the chemical partitioning properties discussed above. More complex global atmospheric and oceanic circulation models have also been used to simulate and forecast the global cycling of chemicals, for example, Guglielmo et al. [17].

Figure 3 shows the variation in long-range transport potential as calculated by the RAIDAR model for a selection of substances.

Bioaccumulation/Bioconcentration and Toxicity

Bioaccumulation refers to the tendency for an organic chemical to partition into biota from their surrounding environment. The bioconcentration factor (BCF) is a metric of bioaccumulation and is the steady-state ratio of the chemical concentration in an organism, such as a fish, to the chemical concentration in an environmental medium, such as water. The BCF is usually measured in a laboratory test and does not include exposures to chemical from dietary sources. Bioaccumulation includes chemical uptake from the environment, that is, bioconcentration, and from food. Biomagnification refers to the increase in chemical concentration from prey to predator [18].

Chemical hydrophobicity, as characterized by K_{OW} , plays a large role in determining the BCF and the toxicity of the chemical as measured in aquatic tests for fish. Hydrophobic organic chemicals bioconcentrate in aquatic organisms as a result of the chemical partitioning from the water into the lipids of the organism. Prolonged bioconcentration of the chemical may induce toxicity by, for example, disrupting the integrity of the lipid bilayer of cell membranes, that is, baseline narcotic mode-of-toxic action. Indeed, the bioaccumulation process results in the transport of chemical from external media, where there are generally no sites for toxic action, into the organism, where there are sites for toxic action [19]. Figure 4 illustrates the fundamental role of partitioning (or bioconcentration) on aquatic toxicity when the parameters for the BCF (L/kg) and toxicity in fish (LC_{50} ; mg/L) are plotted as functions of K_{OW} . Bioconcentration and aquatic toxicity (expressed as $1/LC_{50}$) both increase with K_{OW} and the slopes of the two relationships are almost identical. Chemicals with $\log K_{OW}$ greater than 5 include polycyclic aromatic hydrocarbons (PAHs), many pesticides, chlorinated furans and dioxins, and polychlorinated biphenyls (PCBs) which are implicated as being both toxic and persistent. Chemicals



Toxic Organic Chemicals. Figure 4

Plot of bioconcentration factors (BCF; L/kg, blue circles) in fish and 14-day aquatic toxicity estimates for fish (1/LC50; L/mg, red squares) of selected organic chemicals as a function of chemical hydrophobicity (K_{OW}) [6]

that are highly toxic, such as 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (2,3,7,8-TCDD), and that have modes of action that are more specific than baseline narcosis will typically fall well below the 1/LC50 regression line.

The Stockholm Convention on Persistent Organic Pollutants (POPs)

The Stockholm Convention is an international agreement designed to limit the use of chemicals that are persistent, bioaccumulative, and subject to long-range transport on a continental or global scale such that they cause significant adverse effects (toxicity) in regions distant from where they are used or manufactured [10]. The initial group comprised of 12 substances or classes was colloquially designated the “dirty dozen.” In recent years, other substances have been added. Table 1 lists these substances. It is noteworthy that many are chlorinated hydrocarbons, the reason being that chlorination impacts stability and hence persistence to the molecule.

Chemical Classes

A brief description of some chemical classes is given and provides references for further reading. Molecular structures for selected chemicals in each class are also provided.

Alkanes and Alkenes

The alkanes are present in fuels and generally exhibit low solubilities in water, which when coupled with their relatively high vapor pressures results in relatively high air–water partition coefficients (K_{AW}). Values of K_{OW} are high and increase with carbon number, for example, *n*-hexane has a log K_{OW} value of 4.11 while the higher molecular weight *n*-heptane has a log K_{OW} value of 5.0 [1]. Half-lives in air and water are relatively short (approximately 40 and 200 h in air and water) [1, 3]. The corresponding unsaturated hydrocarbons or alkenes are more soluble in water, have lower K_{AW} and K_{OW} values and are more reactive, especially in air in which half-lives may be only a few hours. They are thus implicated in the formation of photochemical smog due to the presence of the double bond which renders them susceptible to oxidation by hydroxyl radical. This class of chemicals predominantly partitions to air as indicated in Fig. 1.

The halogenated (primarily chlorinated) alkanes are less flammable and are thus valuable solvents but they have longer half-lives and are more environmentally persistent, especially in soils and ground water. They are more hydrophobic than the parent alkanes as a result of their lower solubility in water. Of particular environmental concern are the chlorinated C_2 alkanes and alkenes such as trichloroethylene that are

Toxic Organic Chemicals. Table 1 Chemicals listed under the Stockholm Convention or under review as of 2010; these chemicals are categorized under Annex A (Elimination), Annex B (Restriction), or Annex C (Unintentional production)

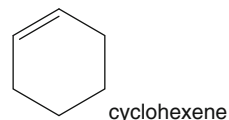
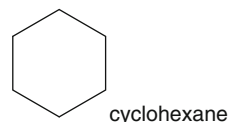
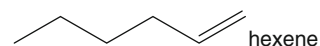
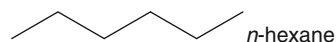
Initial 12 chemicals (i.e., "Dirty Dozen")	
Aldrin (Annex A)	Pesticide
Chlordane (Annex A)	Pesticide
DDT (Annex B)	Pesticide
Dieldrin (Annex A)	Pesticide
Endrin (Annex A)	Pesticide
Heptachlor (Annex A)	Pesticide
Hexachlorobenzene (Annex A, C)	Industrial chemical
Mirex (Annex A)	Pesticide
Toxaphene (Annex A)	Pesticide
Polychlorinated biphenyls (Annex A, C)	Industrial chemical and flame retardant
Polychlorinated dibenzo- <i>p</i> -dioxins (Annex C)	By-product
Polychlorinated dibenzofurans (Annex C)	By-product
Recently listed chemicals	
α -hexachlorocyclohexane (Annex A)	Pesticide, by-product
β -hexachlorocyclohexane (Annex A)	Pesticide, by-product
γ -hexachlorocyclohexane (Annex A)	Pesticide
Chlordecone (Annex A)	Pesticide
Hexabromobiphenyl (Annex A)	Industrial chemical and flame retardant
Pentachlorobenzene (Annex A, C)	Pesticide, industrial chemical
Perfluorooctane sulfonic acid, its salts and perfluorooctane sulfonyl fluoride (Annex B)	
Polybrominated diphenyl ethers (Annex A)	Industrial chemical
Tetra, pentabromodiphenyl ether (Commercial penta)	
hexa- and heptabromodiphenyl ether (Commercial octa)	

Toxic Organic Chemicals. Table 1 (Continued)

Chemicals under review as of 2010	
Endosulfan (Annex F)	Insecticide
Hexabromocyclododecane (Annex E)	Flame retardant
Short-chained chlorinated paraffins (Annex E)	Used in pressure lubricants in metal industry. Also used in sealers, glue coatings in building industry, and in leather and rubber treatments.

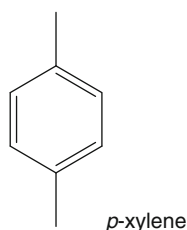
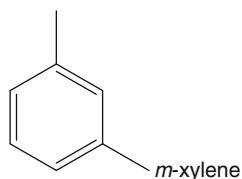
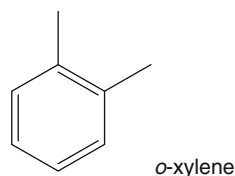
widely used as solvents and have become frequent groundwater contaminants. The chlorinated long- and short-chain paraffin compounds have many uses, notably as cutting oils.

The cycloalkanes are similar in properties to the alkanes. The fully chlorinated cyclohexanes are of concern because of their persistence, volatility, and potential for long-range transport. The pesticide lindane (γ -hexachlorocyclohexane or γ -HCH) and its isomers α -HCH and β -HCH have been distributed globally by transport in air and ocean currents. Uses of these chemicals are now limited. Brominated cyclohexanes that are used as fire retardants are also persistent and bioaccumulative.



Mono-aromatics

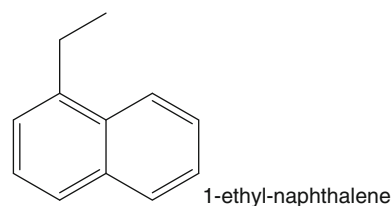
The mono-aromatics, for example, benzene, toluene, and the isomers of xylene, are significant components of fuels and are important chemical intermediates. They have higher solubilities in water than the corresponding cycloalkanes and they are relatively volatile and reactive in air contributing to photochemical smog [3]. Mono-aromatics such as xylene partition mostly to air as indicated in Fig. 1, while higher molecular weight aromatics such as polyaromatic hydrocarbons (e.g., 1-ethyl naphthalene) have lower volatilities and water solubilities. Chlorination results in increased persistence and hydrophobicity (K_{OW}). Hexachlorobenzene is very persistent and has become globally distributed by atmospheric transport and is banned under the Stockholm Convention.



Polyaromatic Hydrocarbons (PAHs)

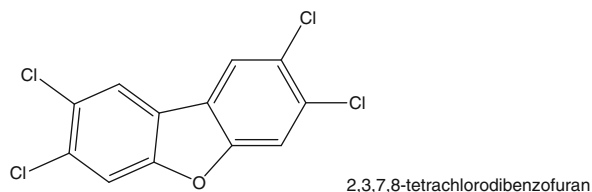
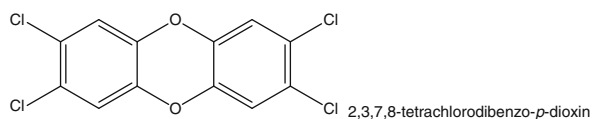
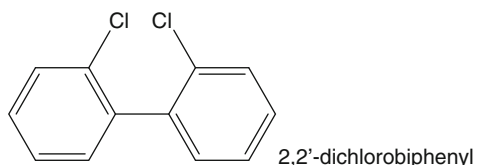
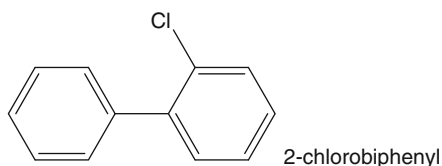
Fusion of a number of benzene rings results in an important series of PAHs such as the 2-ring naphthalene, 3-ring phenanthrene, 4-ring pyrene, and 5-ring benzo-a-pyrene (BaP). These substances are formed along with smoke and soot during incomplete

combustion. They are hydrophobic, relatively involatile, and are regarded as particularly toxic. BaP is a well-studied human carcinogen. The primary environmental concern is human inhalation in the vicinity of combustion sources. Chlorinated naphthalenes are used as PCB substitutes and are hydrophobic and persistent. Chlorination of PAHs increases their hydrophobicity and stability, and decreases their water solubility and volatility. For further reading, see Haritash and Kaushik [20], Mastral and Callen [21], and Wilson and Jones [22].



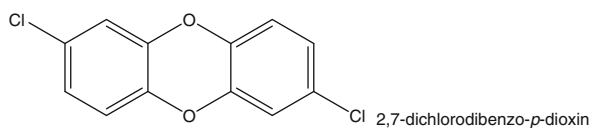
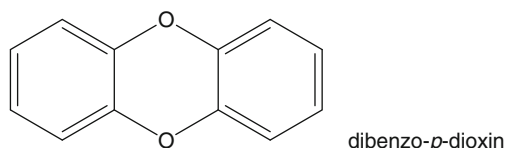
Polychlorinated Biphenyls (PCBs)

Polychlorinated biphenyls (PCBs) are a class of stable synthetic organic chemicals consisting of 209 possible congeners. PCBs are used industrially as dielectric fluids in capacitors and transformers but also as flame retardants, inks, sealants, and plasticizers. The stability of the C-Cl covalent bond makes these chemicals more persistent in the environment. Most biodegradation reactions involve oxidation or reduction of substrate chemicals. PCBs are also resistant to oxidation, making them resistant to biodegradation. As one of the most notorious toxic organics they have been banned under the Stockholm Convention, but large quantities remain in use and they are widespread in distribution. PCBs are associated with adverse health effects in many fish, marine mammal, and avian species. For instance, levels of PCBs above 20 mg/kg in the blubber of whales and seals have been linked to thyroid, immune, and reproductive disruption [23, 24]. PCBs are predicted to partition to octanol more than air and water, but they are sufficiently volatile to be subject to atmospheric transport. For further reading, see also Fox et al. [25], Hoffman et al. [26], Borga et al. [27], Aken et al. [28], Van den Berg et al. [29], and Domingo and Bocio [30].



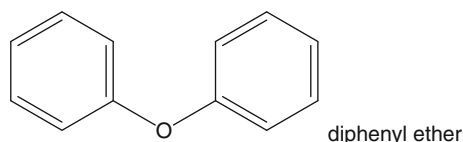
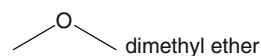
Dioxins and Furans

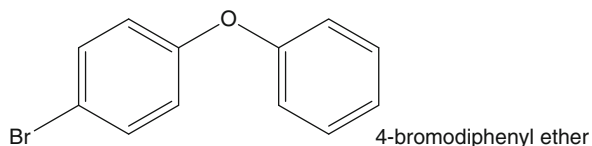
Non-chlorinated dioxins and furans such as dibenzo-dioxin and dibenzo-furan, respectively, are relatively water soluble and volatile as indicated in Fig. 1 and they can partition to all three media. Chlorinated dioxins and furans are by-products of industrial processes and naturally occurring, and are categorized as being unintentionally produced under Annex C of the Stockholm Convention [10]. Chlorination of dioxins and furans results in increased hydrophobicity, lower water solubilities, and lower volatilities, as demonstrated in Fig. 1 where chloro-furans and chloro-dioxins such as 2,8-dichlorodibenzo-furan (2,8-DCDF), 2,7-dichlorodibenzo-dioxin (2,7-DCDD), respectively, are regarded as partitioning largely to octanol, that is, organic phases. Increased chlorination further increases their hydrophobicity as indicated with 2,3,7,8-TCDD and 2,3,7,8-TCDF. A notable congener is 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (2,3,7,8-TCDD), which is a very potent and persistent toxicant, indeed one of the most toxic organic chemicals known. See also Lohmann and Jones [31], Aberg et al. [32], Van den Berg et al. [29], and Domingo and Bocio [30].



Ethers

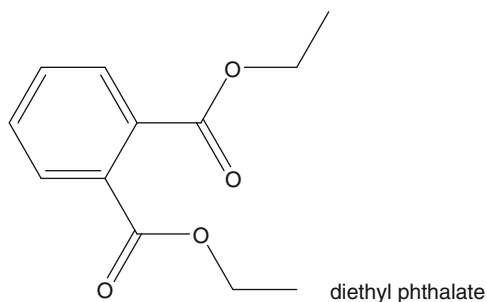
Of more recent environmental concern are certain halogenated ethers such as polybrominated diphenyl ethers (PBDEs). PBDEs are used as fire retardants and are added to a wide variety of commercial and household products. For example, PBDEs are added to polyurethane foam and used in upholstered furniture to reduce their flammability [33]. Commercial pentabromodiphenyl ether, which is mainly composed of tetrabromodiphenyl ether and pentabromodiphenyl ether, is categorized as a persistent organic pollutant and has recently been banned under the Stockholm Convention [10]. Because of their low water solubility, relatively high log K_{OW} , low log K_{AW} , and low vapor pressure, in the environment they partition primarily to sediments and soils. Their persistence in soil and sediments can result in accumulation in the environment and in food webs. Short non-brominated ethers such as dimethyl ether (a common laboratory solvent) are highly soluble, have a high vapor pressure and low log K_{OW} (see Fig. 1), and are less persistent. See also Vonderheide et al. [34], Wang et al. [35], and de Wit et al. [36].





Esters

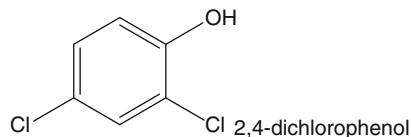
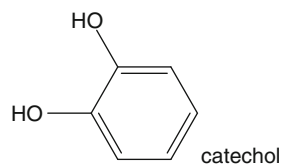
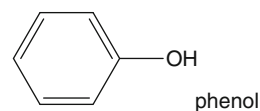
Esters are a product of a condensation reaction between alcohols and carboxylic acids. Esters of low molecular weight are relatively volatile and are components of fragrances. Notably among the esters of environmental concern are the phthalate acid esters (PAEs). These commercial chemicals have been used in a wide range of products namely; household electronics, building materials, insecticides and pharmaceutical products. These compounds are found worldwide in oceans, groundwater, sediment, and the atmosphere [37, 38]. Examples of PAEs include dimethyl phthalate (DMP) and diethyl phthalate (DEP).



Phenols

Phenols are derivatives of benzene possessing a hydroxyl ($-OH$) group. They are used as reagents in the manufacture of dyes, drugs, fungicides, and pesticides and in chemical and paper industries. Sources of emissions of phenolic waste to the environment are from petroleum refineries, industrial plants, mine discharges, and through general uses in products containing resins and paints. The hydroxyl group can dissociate into a phenolate and hydrogen ions imparting acidity and relatively high solubility in water [39]. Substituted phenols such as chlorinated and methylated (2, 4-dichlorophenol, pentachlorophenol (PCP) and 2, 4-dimethylphenol) are more susceptible to dissociation and are more acidic and toxic. PCP has been widely used as a wood

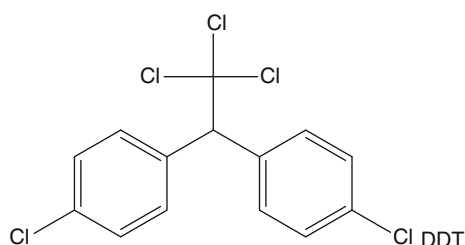
preservative, while phenol was one of the first medical disinfectants. Catechol is a unique phenol with two hydroxyl groups and is implicated in affecting the human nerve center system, inhibits DNA replication, and leads to chromosomal aberration [40]. Derivatives of catechol include the toxic agents in poison ivy and poison oak.



Pesticides

Pesticides are designed to have properties that enable them to reach and impact the target organism or pest including insects, rodents, and vegetation. Effectiveness is increased by persistence, but this can result in nonselective toxicity to other non-targeted organisms including birds and humans. They may have a greater potential to contaminate a wide range of environmental media. Many of the original pesticides were organochlorine (OC) compounds which are now either banned or highly restricted under the Stockholm Convention [10]. Many OCs are relatively volatile; thus, evaporation is an important loss process for pesticides from the areas where they are applied initially, they then move to the atmosphere and eventually get transported through air to remote areas. Notable among the OCs are DDT and its related compounds DDE and DDD, and lindane (γ -HCH). Permethrin, a synthetic pyrethroid insecticide used

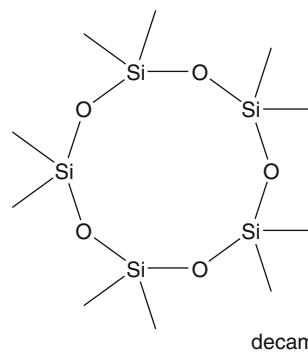
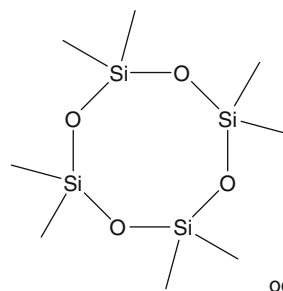
worldwide to control a broad range of insect pests such as mosquitoes, is also highly hydrophobic and partitions in the environment in a manner similar to DDT. OC pesticides have been largely replaced by organophosphates, pyrethroids, and other substances that are less persistent, but often more potent toxicants. For further reading on pesticides, especially on their global distribution, see Weber et al. [41], Hoferkamp et al. [42], Li and Macdonald [43], and Muir and de Wit [44].



Personal Care, Indoor, and Household Products

Most organic substances in personal care and household products can impact humans directly in the home and they may enter the environment through sewage wastes (down the drain releases) or by direct releases to air. Included in this group are detergents, fragrances, flame retardants, additives to plastics, and indoor pesticides and solvents. An example is the class of cyclic siloxanes whose backbone is formed of repeated oxygen and silicon atoms with methyl groups attached. These compounds are components of toiletries such as antiperspirants, soaps, and shampoos and are used in many other “household” products including textiles, paints, sealants, lubricants, and non-medicinal ingredients in pharmaceuticals. They are relatively nontoxic but they can be quite persistent [45]. They are highly volatile with large K_{AW} values and hence predominantly partition to air as illustrated in Fig. 1. Bisphenol A (BPA) (2,2-bis(4-hydroxyphenyl)propane) has been used often in the production of hard plastic (polycarbonate) bottles and metal-based food and beverage can liners in the past few decades. BPA is regarded as an endocrine-disrupting chemical (EDC) [46]. Regulatory agencies have recently expressed serious concerns about the potential effects of BPA on the brain and prostate gland in fetuses, infants, and young

children and some have banned the sale and importation of baby bottles containing BPA [47]. A variety of other organic substances are emitted from various indoor sources, including office equipment such as printers [48].



Conclusions

There is a wide range of organic chemicals that have the potential to be toxic with a multitude of uses and with properties that vary greatly resulting in a diversity of environmental behavior characteristics and exposure routes. Management of these substances presents a considerable challenge to governments and industries. Mistakes have been made in the past by synthesizing and dispersing substances such as DDT and PCBs that have proven to be persistent, bioaccumulative, toxic and have become globally distributed. There is optimism that such mistakes can be avoided in the future if there is an adequate understanding of the factors that dictate environmental partitioning, fate, exposure, and toxicity. Effort is now being devoted to searching for new toxic organics by advance knowledge of their

structural and property characteristics [49–51]. It is increasingly possible to estimate the critical chemical properties from molecular structure and assess environmental fate and transport quantitatively using mass balance models that can be validated using monitoring data. If this information can be exploited in advance of commercial distribution, it should be possible to enjoy the many benefits of synthetic organic chemicals without the adverse effects that have been, and are being experienced, from uses that do not satisfy the principles of conservation and sustainability.

Future Directions

The effective future management of toxic organic chemicals in the environment can be improved by a number of scientific and regulatory actions. There is a need for better information on the production, uses and rates of discharge of chemicals, both nationally and internationally. Improved methods of estimating chemical properties from molecular structure are desirable including partitioning, degradability, and toxicity. These improvements could enhance the assessment of chemical fate and possible effects by industry, thus avoiding repetition of past mistakes. Undoubtedly, there are as yet unidentified toxic organic substances, thus there is an incentive to improve techniques and sensitivity of chemical analyses and monitoring of environmental media. As a complement to these monitoring efforts there is a need to further develop, apply, and validate the use of computer models that predict environmental fate and effects. Finally there is a strong incentive to increase the effectiveness of regulatory programs, both nationally and internationally by incorporating the emerging scientific advances and providing a holistic and transparent evaluation of risk to the public and the ecosystem. Ultimately it is an informed public that is the best guarantee that society will maintain and enhance actions to improve sustainability and conservation.

Bibliography

1. Mackay D, Shiu WY, Ma KC (2000) Physical-chemical Properties and Environmental Fate Handbook. Chapman and Hall CRCnetBASE, CRC Press LLC, Boca Raton
2. Seth R, Mackay D, Muncke J (1999) Estimating the organic carbon partition coefficient and its variability for hydrophobic chemicals. *Environ Sci Technol* 33(14): 2390–2394
3. Atkinson R, Arey J, Aschmann SM (2008) Atmospheric chemistry of alkanes: review and recent developments. *Atmos Environ* 42(23):5859–5871
4. Howard PH, Boethling RS, Jarvis WF, Meylan WM, Michalenko EM (1991) Handbook of Environmental Degradation Rates. Lewis Publishers, Chelsea
5. Arnot JA, Mackay D, Parkerton TF, Bonnell M (2008) A database of fish biotransformation rates for organic chemicals. *Environ Toxicol Chem* 27(11):2263–2270
6. U.S. EPA (2009) Exposure assessment tools and models, estimation programs interface (EPI) suite, ver. 4.0. U.S. Environmental Protection Agency, Exposure Assessment Branch, Washington, DC
7. Government of Canada, Persistence and Bioaccumulation Regulations. Canada Gazette Part II March 29, 2000
8. Webster E, Mackay D, Wania F (1998) Evaluating environmental persistence. *Environ Toxicol Chem* 17:2148–2158
9. Wegmann F, Cavin L, MacLeod M, Scheringer M, Hungerbühler K (2009) The OECD software tool for screening chemicals for persistence and long-range transport potential. *Environ Modell Softw* 24(2):228–237
10. UNEP (2001) Final Act of the Conference of Plenipotentiaries on the Stockholm Convention on Persistent Organic Pollutants. In: United Nations Environment Program, Geneva, Switzerland, p 44
11. Gouin T, Mackay D, Webster E, Wania F (2000) Screening chemicals for persistence in the environment. *Environ Sci Technol* 34(5):881–884
12. Mackay D, McCarthy LS, MacLeod M (2001) On the validity of classifying chemicals for persistence, bioaccumulation, toxicity and potential for long-range transport. *Environ Toxicol Chem* 20(7):1491–1498
13. Scheringer M (2002) Persistence and Spatial Range of Environmental Chemicals. Wiley-VCH, Weinheim
14. Wania F (2003) Assessing the potential of persistent organic chemicals for long-range transport and accumulation in polar regions. *Environ Sci Technol* 37(7):1344–1351
15. Wania F, Mackay D (1995) A global distribution model for persistent organic-chemicals. *Sci Total Environ* 161: 211–232
16. Arnot JA, Mackay D, Webster E, Southwood JM (2006) Screening level risk assessment model for chemical fate and effects in the environment. *Environ Sci Technol* 40(7):2316–2323
17. Guglielmo F, Lammel G, Maier-Reimer E (2009) Global environmental cycling of γ -HCH and DDT in the 1980s – a study using a coupled atmosphere and ocean general circulation model. *Chemosphere* 76(11):1509–1517
18. Gobas FAPC, Morrison HA (2000) Bioconcentration and Biomagnification in the Aquatic Environment. In: Boethling RS, Mackay D (eds) Handbook of Property Estimation Methods for Chemicals: Environmental and Health Sciences. CRC Press, Boca Raton

19. McCarty LS, Mackay D (1993) Enhancing ecotoxicological modeling and assessment: body residues and modes of toxic action. *Environ Sci Technol* 27(9):1719–1728
20. Haritash AK, Kaushik CP (2009) Biodegradation aspects of polycyclic aromatic hydrocarbons (PAHs): a review. *J Hazard Mater* 169(1–3):1–15
21. Mastral AM, Callen MS (2000) A review on polycyclic aromatic hydrocarbon (PAH) emissions from energy generation. *Environ Sci Technol* 34(15):3051–3057
22. Wilson SC, Jones KC (1993) Bioremediation of soil contaminated with polynuclear aromatic hydrocarbons (PAHs): a review. *Environ Pollut* 81(3):229–249
23. Kannan K, Blankenship AL, Jones PD, Giesy JP (2000) Toxicity reference values for the toxic effects of polychlorinated biphenyls to aquatic mammals. *Hum Ecol Risk Assess* 6:181–201
24. Hornbuckle K, Robertson L (2010) Polychlorinated Biphenyls (PCBs): sources, exposures, toxicities. *Environ Sci Technol* 44(8):2749–2751
25. Fox GA, Weseloh DV, Kubiak TJ, Erdman TC (1991) Reproductive outcomes in colonial fish-eating birds – a biomarker for developmental toxicants in Great-Lakes food-chains. 1. Historical and ecotoxicological perspectives. *J Great Lakes Res* 17(2):153–157
26. Hoffman DJ, Rice CP, Kubiak TJ (1996) PCBs and dioxins in birds. In: Beyer WN, Heinz GH, Redmon-Norwood AW (eds) *Environmental Contaminants in Wildlife: Interpreting Tissue Concentrations*. Lewis Publishers, Boca Raton, pp 165–207
27. Borga K, Fisk AT, Hargrave B, Hoekstra PF, Swackhamer D, Muir DCG (2005) Bioaccumulation factors for PCBs revisited. *Environ Sci Technol* 39(12):4523–4532
28. Aken BV, Correa PA, Schnoor JL (2009) Phytoremediation of polychlorinated biphenyls: new trends and promises. *Environ Sci Technol* 44(8):2767–2776
29. Van den Berg M, Birnbaum LS, Bosveld ATC, Brunstrom B, Cook P, Feeley M, Giesy JP, Hanberg A, Hasegawa R, Kennedy SW, Kubiak T, Larsen JC, van Leeuwen FXR, Liem AKD, Nolt C, Peterson RE, Poellinger L, Safe S, Schrenk D, Tillitt DE, Tysklind M, Younes M, Waern F, Zacharewski T (1998) Toxic equivalency factors (TEFs) for PCBs, PCDDs PCDFs for humans and wildlife. *Environ Health Perspect* 106(12):775–792
30. Domingo JL, Bocio A (2007) Levels of PCDD/PCDFs and PCBs in edible marine species and human intake: a literature review. *Environ Int* 33(3):397–405
31. Lohmann R, Jones KC (1998) Dioxins and furans in air and deposition – a review of levels, behavior and process. *Sci Total Environ* 219:53–81
32. Aberg A, MacLeod M, Wiberg K (2008) Physical-chemical property data for dibenzo-p-dioxin (DD), dibenzofuran (DF), and chlorinated DD/Fs: a critical review and recommended values. *J Phys Chem Ref Data* 37(4):1997–2008
33. Hites RA (2004) Polybrominated diphenyl ethers in the environment and in people: a meta-analysis of concentrations. *Environ Sci Technol* 38(4):945–956
34. Vonderheide AP, Mueller KE, Meija J, Welsh GL (2008) Polybrominated diphenyl ethers: causes for concern and knowledge gaps regarding environmental distribution, fate and toxicity. *Sci Total Environ* 400(1–3):425–436
35. Wang Y, Jiang G, Lam PKS, Li A (2007) Polybrominated diphenyl ethers in the East Asian environment: a critical review. *Environ Int* 33(7):963–973
36. de Wit CA, Herzke D, Vorkamp K (2010) Brominated flame retardants in the Arctic environment – trends and new candidates. *Sci Total Environ* 408(15):2885–2918
37. Staples CA, Peterson DR, Parkerton TF, Adams WJ (1997) The environmental fate of phthalate esters: a literature review. *Chemosphere* 35(4):667–749
38. Peijnenburg WJGM, Struis J (2006) Occurrence of phthalate esters in the environment of the Netherlands. *Ecotoxicol Environ Saf* 63:204–215
39. Shiu W, Ma K-C, Varhanickova D, Mackay D (1994) Chlorophenols and alkylphenols: a review and correlation of environmentally relevant properties and fate in an evaluative environment. *Chemosphere* 29(6):1155–1224
40. Topping DC, Bernard LG, O'Donoghue JL, English JC (2007) Hydroquinone: acute and subchronic toxicity studies with emphasis on neurobehavioral and nephrotoxic effects. *Food Chem Toxicol* 45:70–78
41. Weber J, Halsall CJ, Muir D, Teixeira C, Small J, Solomon K, Hermanson M, Bidleman T (2010) Endosulfan, a global pesticide: a review of its fate in the environment and occurrence in the Arctic. *Sci Total Environ* 408: 2966–2984
42. Hoferkamp L, Hermanson MH, Muir DCG (2010) Current use pesticides in Arctic media; 2000–2007. *Sci Total Environ* 408(15):2985–2994
43. Li YF, Macdonald RW (2005) Sources and pathways of selected organochlorine pesticides to the Arctic and the effect of pathway divergence on trends in biota: a review. *Sci Total Environ* 342(1–3):87–106
44. Muir DCG, de Wit CA (2010) Trends of legacy and new persistent organic pollutants in the circumpolar arctic: overview, conclusions, and recommendations. *Sci Total Environ* 408:3044–3051
45. Graiver D, Farminer KW, Narayan R (2003) A review of the fate and effects of silicones in the environment. *J Polym Environ* 11(4):129–136
46. Hiroi T, Okada K, Imaoka S, Osada M, Funae Y (2006) Bisphenol A binds to protein disulfide isomerase and inhibits its enzymatic and hormone-binding activities. *Endocrinology* 147:2773–2780
47. Meeker JD, Calafat AM, Hauser R (2010) Urinary bisphenol A concentrations in relation to serum thyroid and reproductive hormone levels in men from an infertility clinic. *Environ Sci Technol* 44(4):1458–1463
48. Destailhats H, Maddalena RL, Singer BC, Hodgson AT, McKone TE (2008) Indoor pollutants emitted by office equipment: a review of reported data and information needs. *Atmos Environ* 42(7):1371–1388

49. Brown TN, Wania F (2008) Screening chemicals for the potential to be persistent organic pollutants: a case study of arctic contaminants. *Environ Sci Technol* 42(14): 5202–5209
50. Howard PH, Muir DCG (2010) Identifying new persistent and bioaccumulative organics among chemicals in commerce. *Environ Sci Technol* 44(7):2277–2285
51. Muir DCG, Howard PH (2006) Are there other persistent organic pollutants? A challenge for environmental chemists. *Environ Sci Technol* 40(23):7157–7166

Transgene Expression in Plants, Control of

SYLVIE DE BUCK¹, ANNELIES DE PAEPE², ANN DEPICKER³

¹Department of Plant Biotechnology and Genetics, Ghent University, Ghent, Belgium

²Department of Plant Systems Biology, Flanders Institute for Biotechnology VIB, Ghent, Belgium

³UGent Vakgroep Plantenbiotechnologie en Genetica, VIB Plant Systems Biology Department, Ghent University, Ghent, Belgium

Article Outline

Glossary

Definition of the Subject

Introduction

Generation of Transgenic Plants by *Agrobacterium*-Mediated Transformation or Direct Gene Transfer

Transgene Loci Can Be Either Simple or Complex Regulatory Sequences Have a Major Impact on

Transgene Expression and Expression Variability The Influence of the Integration Position on Transgene Expression

Correlation Between T-DNA Locus Structure and Homology-Based Silencing

Control of Gene Expression by the Generation of Single-Copy Transformants

Use of Gene Silencing Mutants to Overcome Transgene Expression Variability

Future Directions

Acknowledgments

Bibliography

Glossary

CRE/*loxP* A site-specific recombination system in which the CRE recombinase catalyzes recombination between *loxP* sequences. A *loxP* sequence is 34 bp long and consists of two inverted repeats of 13 bp and one spacer region of 8 bp. The orientation of the *loxP* sequence is determined by the spacer region.

Inverted T-DNA repeat Repeat obtained from the integration of two T-DNA copies at one genetic locus, but with one T-DNA integrated in inverted orientation compared to the other.

Position effect Influence of the position of a gene in the genome upon its expression.

Posttranscriptional gene silencing (PTGS) Silencing mechanism leading to the sequence-specific degradation of target mRNAs and sequence-specific suppression of translation.

Single-copy transformant Transgenic plant harboring only one copy of the introduced DNA segment.

Transcriptional gene silencing (TGS) Silencing mechanism that targets homologous DNA sequences in the promoter, suppressing transcription and correlating with promoter sequence methylation.

Transfer DNA (T-DNA) The DNA fragment that is delineated by the right and left border repeats and is transferred from *Agrobacterium* to the plant cell by the type-IV secretion system.

Transgene expression variability Variability that is higher in a population of transformants than expected based on gene dosage effects.

Transgenic plant Plant harboring one or more external DNA segments that had been introduced and stabilized by integration into the plant genome. The foreign DNA is transferred to the plant cell via *Agrobacterium*-mediated transformation or direct gene transfer.

Definition of the Subject

To define the subject in relation to the title, we would like to emphasize that most of the results and conclusions described below were obtained in the model plants *Arabidopsis thaliana* and *Nicotiana tabacum* after *Agrobacterium*-mediated transformation, and in

a limited number of crops, such as wheat, rice, and soybean, after direct gene transfer. Additionally, most of the transgenes described below are transcriptional fusions with the 35S promoter from the cauliflower mosaic virus. Indeed, to obtain transgenic plants with high and constitutive expression of a transgene, the 35S promoter is commonly used. This 35S promoter is very powerful, because it generally leads to constitutively high transcript and protein levels of the transgene in most dicot plants and is not greatly influenced by environmental conditions or tissue types. However, over the years, it became clear that 35S-driven transgene expression is very variable in the different transformants of a transgenic plant population with the same construct. For most analyzed overexpression constructs containing this 35S promoter, more than a 100-fold difference in protein accumulation levels has been observed. Approximately 20% of the transformants have high recombinant protein accumulation levels, but approximately 80% display an intermediate or low and unstable transgene expression. Additionally, high-expression transformants often segregate progeny plants with low recombinant protein accumulation levels. Thus, it is imperative that for scientific analysis and applications, the variability of 35S-driven transgenes is known in different transformants and generations. In this contribution, the possible causes of variation in transgene expression in a population of transgenic plants are discussed and several approaches to diminish the variability are reviewed.

Introduction

From the mid-1980s on, transgenic plants were generated by adding one or more genes to a plant's genome. Transformation was usually achieved with gold-particle bombardment or through the gene transfer process via *Agrobacterium tumefaciens*, a soil bacterium carrying an engineered T-DNA vector. The inserted gene sequence, known as the transgene, can be derived from any living prokaryotic or eukaryotic organism. Transgenic plants were originally engineered to address specific scientific questions and to investigate the plant biology in general. However, the ability to create transgenic plants also opened the way to manipulate crop plants for better characteristics, such as increased yield, enhanced quality, pest or disease resistance, superior

tolerance to heat, cold, and drought, and nutritional improvement. In addition, transformation of plants allows molecular farming of many limiting components, such as proteins, metabolites, fatty acids, and bioenergy precursors.

To generate transgenic plants, transformation methods had to be explored and ameliorated. Furthermore, the vectors and expression cassettes for controlled integration and expression of the introduced transgenes had to be worked out. Once the *Agrobacterium*-based system for T-DNA transfer had been unraveled, the simple view was that a particular T-DNA sequence could be constructed to be integrated as a single copy into the plant genome. However, it turned out that transformants quite often integrate two to ten T-DNA copies at one locus or dispersed over several loci [1]. Moreover, the transferred DNA is integrated randomly and, thus, every transformant containing the same transgene differs regarding the integration site and the transgene copy number [2]. Concerning transgene expression, in principle, it is very straightforward to combine particular gene-controlling elements with a coding sequence of interest to obtain a predictable level of transgene expression. The most frequently used promoter elements are the promoter of the nopaline synthase gene (pNOS), driving weak and constitutive expression of the downstream transcribed sequence, and the promoter of the cauliflower virus 35S transcript (p35S), driving in general at least 20-fold higher expression levels. Whereas the majority of transformants with transgenes controlled by the pNOS promoter display the same mRNA level, independent transgenic lines generated with a p35S-controlled transgene construct show variable transgene expression with more than 100-fold difference in recombinant protein levels. In the latter case, frequently the recombinant protein accumulation levels in the different transformants are distributed bimodally [3]. Low p35S transgene expression occurs regularly in transgenic lines with multiple T-DNA copies, particularly when they are integrated into inverted orientation. As a consequence, the transgenes are transcribed convergently [4]. However, also single T-DNA copies can result in low or even no transgene expression [5]. Originally, the mean of recombinant protein accumulation levels in different transformants was considered to represent the expression strength of a particular

transgene construct. Afterward, it was realized that the class of transformants with high transgene expression correlated better with the transgene expression capacity [6], implying that all transformants with intermediate and low transgene expression have controlling mechanisms that result in transgene suppression or transgene silencing.

A first and visually dramatic example of transgene silencing was found in plants that overexpressed gene constructs coding for pigment production in wild-type petunia (*Petunia hybrida*) with purple flowers [7, 8]. Contrary to their expectations, the flowers of the transgenic petunia exhibited reduced and variable instead of increased pigmentation. As both the expression of endogenes and transgenes were suppressed, this phenomenon was called cosuppression [7, 8]. Later on, the transgene silencing mechanisms were divided into three groups [9]. One group harbors the transgene silencing events based on position effects and heterochromatinization. As transgenes integrate randomly, different transformants contain the transgene at a different chromosomal position and nearby located chromatin-controlling elements, enhancers and silencers, are expected to influence the activity of the integrated gene. The other two groups of silencing mechanisms operate via homology between the silencer sequences and the target silenced genes, do not depend on position, and exert their effects *in trans* [9, 10]. Transcriptional silencing (TGS) refers to the silencing process that suppresses transcription of the silenced genes based on homology in the silencer locus-containing promoter region. TGS functions at the level of prevention of transcription initiation of the silenced genes; is often accompanied with concomitant DNA methylation, altered histone modifications, and heterochromatin formation; and is stably inherited through meiosis and mitosis [11]. In contrast, post-transcriptional gene silencing (PTGS) needs homology in the transcribed region, results in specific transcript degradation and/or specific repression of translational initiation, and resets through sexual propagation [10, 12].

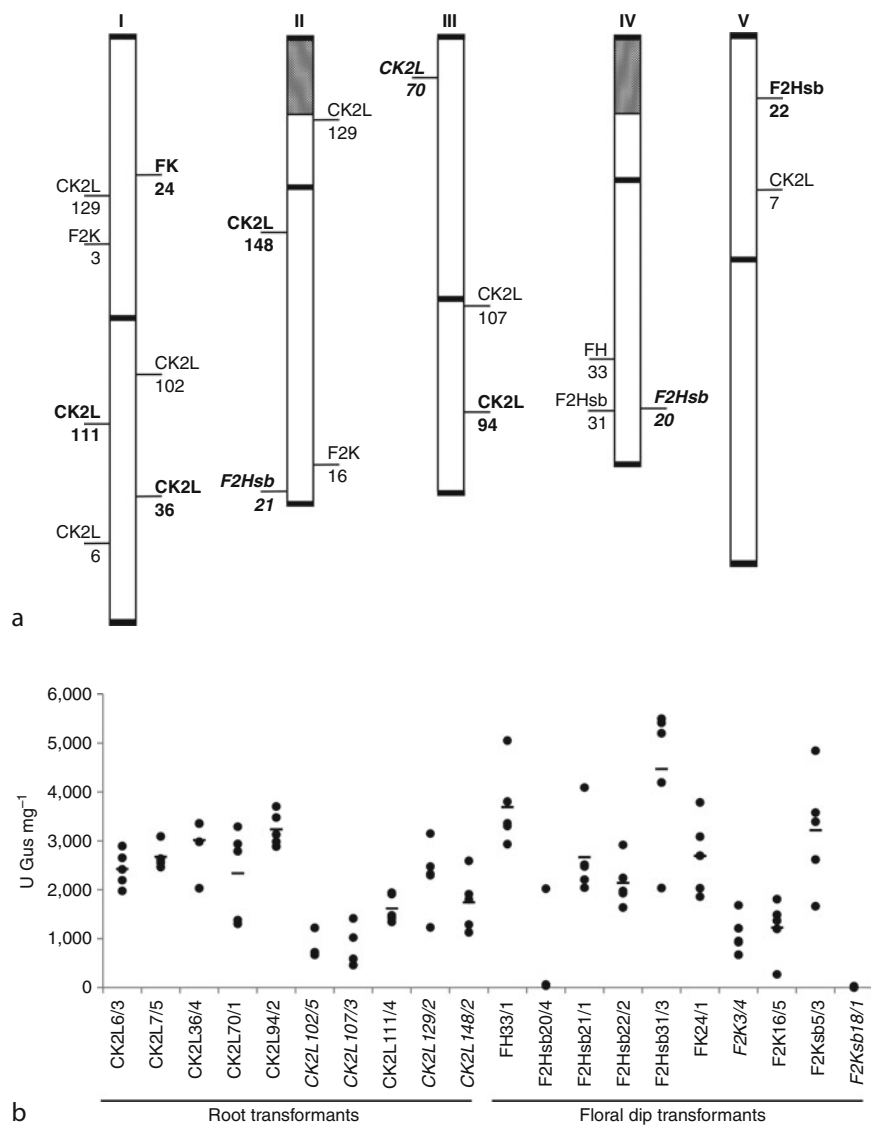
Here, different parameters that influence transgene expression in plants are described: (1) the transformation process, which results in plants with transgene loci that can be either simple or complex; (2) the different regulatory sequences used in the transgene expression

construct; (3) the position of the integrated transgene into the plant genome; and (4) the correlation between repeat structures and transgene silencing. Several approaches to minimize the transgene expression variability are discussed: (1) the resolution of complex integration loci by the CRE/*loxP* recombination process, (2) site-specific or site-directed integration, and (3) the use of gene silencing mutants to create a population of transgenic plants with high and stable transgene expression.

Generation of Transgenic Plants by *Agrobacterium*-Mediated Transformation or Direct Gene Transfer

Agrobacterium-mediated transformation is the most widely used method to generate transgenic plants [13, 14], although direct transformation methods are still very important for certain crop plants (see below). During *Agrobacterium* transformation, the transferred DNA (T-DNA), delineated by the right border (RB) and left border (LB) repeats, is integrated into the plant genome by illegitimate recombination. Overall, approximately one half of the transformants contains only one T-DNA copy and the other half multiple T-DNAs (two to ten), integrated at one or several independent loci [2, 15].

The T-DNA integrates at random in any chromosome [2] (Fig. 1) and is as efficient in all five chromosomes of *Arabidopsis thaliana* as in intergenic and intragenic sequences and in exons and introns (Fig. 1) [16, 17]. With promoterless selectable markers and consecutive selection on those markers, many transformants could be recovered, indicating that T-DNAs frequently integrate in transcriptionally active regions [18, 19], but in the absence of transcriptional activation selection, many transformants integrated the T-DNA between genes and in repetitive DNA. Under nonselective conditions, silent T-DNA insertions were found in heterochromatic regions, centromeres, telomeres, and rDNA repeats [20, 21]. T-DNA integration is accompanied with small deletions of the plant target DNA and/or T-DNA ends. Very often, microhomology of 2–10 bp is observed between the plant pre-insertion site and the T-DNA ends, suggesting that T-DNA integration occurs via the nonhomology-dependent double-stranded break repair [2, 15, 22].



Transgene Expression in Plants, Control of. Figure 1

Random integration of single-copy T-DNAs into the *Arabidopsis* genome and the similar GUS activity levels yielded by single-copy 35S-*GUS* transgenes in the different transformants (Adapted from [17]). (a) Distribution of 18 single-copy T-DNA transformants on the five chromosomes of the *Arabidopsis* genome. In nine transformants, the T-DNA is integrated into an intergenic region and nine in a transcribed annotated gene (indicated in *bold*). Of these nine intragenic inserted T-DNAs, six were integrated into an exon and three into an intron (indicated in *italics*). Centromeres and telomeres are indicated in *vertical black boxes* and rDNA repeats in *gray boxes*. For transformant CK2L129, the LB and RB regions were fused to sequences of chromosomes 1 and 2. (b) GUS activity analysis in the T2 generation of 20 different single-copy T-DNA transformants. GUS activity was measured in the leaves of five 6-week-old seedlings per transformant (represented as a *dot*). All transformants were homozygous for the T-DNA insertion, except for the transformants indicated in *italics*. The average GUS activity level in the five seedlings of the homozygous transformants is marked by a *line*. The intra-transformant variability in GUS activity was as high as the inter-transformant variability, except for two transformants without detectable GUS activity. GUS activity levels are given as units GUS mg⁻¹ of total soluble protein. For transformants F2Ksb5 and F2Ksb18, the integration position into the genome could not be determined (see a)

In theory, only the T-DNA, delineated by the LB and RB, is transferred to the plant cell, but in many plants, vector backbone sequences are transferred and integrated into the genome as well [23–26]. Two mechanisms might account for this vector backbone transfer. First, the LB repeat could mistakenly be recognized as initiation site for T-strand production and, as such, result in the transfer of vector backbone sequences. Second, the consequence of inefficient recognition of the LB repeat might be read-through from the T-DNA into the vector sequences [23, 27]. Both mechanisms can be distinguished by analyzing whether or not the LB T-DNA sequences are directly linked to the vector sequences. The vector backbone transfer and integration is seemingly not influenced by the plant species, the explant type used for transformation, the T-DNA vector replicon type, or the selection [25]. Molecular DNA blot and polymerase chain reaction analyses revealed that the vector backbone sequences were mostly linked to both the LB and RB T-DNAs. In fact, the complete vector backbone sequence was integrated between two in tandem oriented T-DNAs, emphasizing the importance of the second mechanism in which the integration of complete vector backbone sequences results from a conjugative transfer initiated at the RB, followed by copying of the T-DNA, the vector, and again a T-DNA from a circular template due to read-through of consecutively the LB and RB [25]. Furthermore, the results demonstrated that neither the RB nor the LB was efficiently recognized as termination and initiation sites, respectively [25]. Additional evidence that the surrounding regions of the border repeats are important for the recognition of the repeat as initiation or termination site was obtained from hybrid border regions. In the absence of the surrounding border regions, the LB repeat was not recognized as T-DNA termination site; addition of the natural LB inner region and occurrence of both the octopine and nopaline LB regions with their repeats, improved the correct recognition of the LB repeat when compared to an LB with only the LB outer region [26].

A number of agronomically important plant species are still recalcitrant for *Agrobacterium*-mediated transformation and are, therefore, transformed with direct gene transfer methods, such as particle bombardment or biolistics, electroporation, and polyethylene glycol transformation [28]. A disadvantage of these

techniques is that transformants tend to have multiple transgene copies integrated as a concatemeric array at one locus, and contain inverted repeats [29–33]. However, several approaches have been described to circumvent this problem (see below).

Transgene Loci Can Be Either Simple or Complex

Upon *Agrobacterium*-mediated transformation, a significant number of transformants contain a single-copy T-DNA insert, but, dependent on the transformation method used, a high percentage of transformants contain multiple T-DNAs integrated into the plant host genome. These multiple T-DNA copies are mostly clustered in one genetic locus, but they can also be present in two or more loci. Multiple T-DNAs integrated at one genetic locus are organized as direct and/or inverted repeats of the T-DNA segment. Complex T-DNA loci were found in many plant species, such as *Arabidopsis*, tobacco (*Nicotiana tabacum*), petunia, and potato (*Solanum tuberosum*) [34–38]. As direct and inverted T-DNA repeats were not found in the bacteria, they are believed to be formed in the plant cell prior to or during the T-DNA integration [34, 35, 39, 40].

Two models have been proposed for the formation of multicopy T-DNA loci upon *Agrobacterium*-mediated transformation: the replication and the ligation models [38]. In the replication model, the repeats originate from a single T-DNA copy that is replicated after introduction in the plant cell and before or during integration into the genome. This hypothesis is favored because all T-DNAs involved in repeat structures had analogous breakpoints in a restriction analysis and because three identical T-DNA copies were observed after transformation with a library of T-DNAs containing different promoters [38, 41, 42]. However, cotransformation experiments with different T-DNAs originating from different *agrobacteria* gave, with a similar frequency, rise to T-DNA loci with different T-DNAs in direct or inverted orientation [1, 34, 35, 43–46]. As these T-DNA structures cannot be formed by replication, but only by ligation of the cotransformed T-DNA copies before or during integration [34, 35], the ligation model postulates that repeats originate from extrachromosomal ligation of two or more individual T-DNAs prior to or during integration into the genome [34] and even that repeat structures are formed by the cointegration of

several T-strand intermediates in one target site rather than by ligation of T-strands [39]. In any case, T-DNA inverted repeats about the RB, an integration structure that is frequently observed upon *Agrobacterium*-mediated transformation, can only be formed after duplication of the transferred single-stranded T-DNAs. Sequence analysis of T-DNA junctions revealed that, during the formation of T-DNA repeat structures, end-to-end ligation of double-stranded T-DNAs occurs especially between right T-DNA ends, whereas recombination based on microhomology regions and insertions of filler DNA was more frequent at LB junctions [22, 35]. None of the sequenced T-DNA/T-DNA junctions contained plant DNA, suggesting that T-DNA recombination and ligation occurred in the majority of cases before integration [35].

Analysis of which parameter determined the structure and complexity of the transgene locus indicated a possible correlation with either the used *Agrobacterium* strain, the titer, or the physiology of the bacterial inocula, bacterial vectors, plant species, or ecotype, but overall no clear picture was obtained [1, 34, 38, 44, 47, 48]. Single-copy T-DNA insertions were found predominantly upon *Arabidopsis* root transformation, whereas multiple insertions, especially organized in T-DNA repeat structures over the RB occurred in transformants generated by leaf disc transformation [48]. However, this was not found in root and leaf disc transformants in an independent experiment (M. De Neve and A. Depicker, unpublished results), indicating that other parameters had to be involved. Nevertheless, a detailed study revealed that the structure of the T-DNA integration locus is especially determined by the transformed target cell [1, 48]. Whereas 20 years ago, *Agrobacterium*-mediated transformation of *Arabidopsis* was achieved mainly by in vitro tissue explant methods, such as root transformation [49], nowadays floral dip is commonly used [50], but the obtained T-DNA copy number in transformants obtained after both transformation methods varies much. To analyze whether this difference was due to cotransformation frequencies or to replication in the plant cell, several floral dip and root transformations were done with mixtures of *Agrobacterium* strains, each carrying one or two different T-DNA vectors, allowing to trace back the origin of complex T-DNA loci [1]. Although the cotransformation frequencies of

T-DNAs originating from different *agrobacteria* after floral dip and root transformation were comparable, and the frequencies of T-DNA originating from one bacterium only slightly higher than those after floral dip transformation, the T-DNA copy numbers in the transformants differed completely. Upon floral dip transformation, on average T-DNA copies were integrated at one genetic locus versus one to three after root transformation [1]. The cotransformation frequencies might explain the T-DNA copy number in most root transformants, but not in floral dip transformants. Therefore, it was postulated that T-DNA replication of a single T-strand occurs in the plant cell before or during T-DNA integration and that its frequency is much higher upon floral dip than upon root transformation [1]. Because the same *Agrobacterium* strains had been used in both transformation methods, not the bacterium but the type of target cell determined the complexity of the T-DNA pattern [1].

In general, transformants obtained after direct gene transfer contain more transgene copies than after *Agrobacterium*-mediated transformation [30–32]. In maize (*Zea mays*), comparison of transgene copy numbers and RNA expression levels upon transformation via direct gene transfer and *Agrobacterium*-mediated transformation revealed that more than 90% of the transformants obtained after *Agrobacterium*-mediated transformation harbored fewer than three T-DNA copies, while most of the transformants obtained after particle bombardment contained more than three copies [32], of which some of the transformants contained as many as 100 copies of the transgene [32]. Another observation was that the transformants obtained after *Agrobacterium*-mediated transformation displayed a higher transgene expression than after particle bombardment. Besides the expected positive correlation between transgene copy number and transgene expression, frequently a negative correlation between gene dosage and transgene expression was observed (see below) [51].

Recent research revealed that more single-copy integrants could be obtained by particle bombardment when a limited quantity of DNA was used for bombardment [52, 53]. A decrease in DNA from 1.5 to 2.5 ng per shot resulted in an increase of single-copy transformants from 30% to 70% [52, 53]. The 10% reduction in transformation efficiency was more than

compensated by the more efficient screening for single-copy transformants [53]. Additionally, complex integration patterns could also be avoided when cassette DNA, harboring a promoter, a coding region and a 3' end region, instead of whole plasmids were used for bombardment [52, 54–56].

Regulatory Sequences Have a Major Impact on Transgene Expression and Expression Variability

To obtain transformants with high and stable transgene expression, the construct should be designed carefully because several sequences might have a great impact on the transgene expression and expression variability. Especially, the selected promoter that drives the transgene expression will determine the accumulation levels of the transgene transcript and, indirectly, of the transgene-encoded protein. Since the beginning of chimeric gene constructions, plant scientists predominantly chose the 35S promoter of the cauliflower mosaic virus as a strong and constitutive promoter to overexpress a coding sequence of interest [57] and for the promoters of the nopaline (pNOS) and octopine (pOCS) synthase-encoding genes, as weak promoters, for instance to drive the selectable marker genes in dicots.

The strength of the 35S promoter increases when upstream activator sequences are added [58], but as a consequence, single-copy transgenes driven by a 35S promoter with two upstream activator regions are more prone to gene silencing [59–61]. In general, silencing frequency correlates positively with the promoter strength [61–64]. Indeed, silencing did not occur frequently when a weak pNOS promoter was used [62], whereas the 35S promoter is very prone to transgene silencing, most probably because the high expression leads to transgene RNA accumulation levels above a certain threshold that triggers RNA degradation [65]. Strikingly, promoters do not only influence the levels, but also the variability of expression in a population of independently transformed plants with the same transgene [66]. Transformation of *Arabidopsis* plants with a p35S- β -glucuronidase (*GUS*) construct resulted in a bimodal expression pattern, with more than 80% of the primary transformants showing a very low and less than 20% a high *GUS* accumulation [3, 66]. The mannopine synthase

promoter (pMAS) yielded gene expression levels that were only slightly lower than those of the 35S promoter, but the variation in transgene expression was at least eightfold lower [66]. The bimodal character of the p35S promoter versus the normal distribution of the pMAS-driven gene expression suggests that silencing phenomena occur less frequently in plants transformed with pMAS than with p35S.

Not only the promoter, but also the codon usage might play a role in determining the transcript stability and/or translation efficiency. For *Arabidopsis*, highly expressed endogenous genes seem to contain a C at the third codon position, whereas genes with low expression levels have predominantly a T in that position [67]. By adapting codon usage in the green fluorescent protein (GFP)-encoding gene toward a higher proportion of codons with a C or a G in the third position, more highly *GFP*-expressing plants were generated [68]. Additionally, the nature of the coding sequence and the sequence composition seem to be important parameters in determining the transcript accumulation and/or threshold above which silencing is initiated [62, 69]. The transcript levels of the three reporter genes *GUS*, *GFP*, and streptomycin phosphotransferase (*SPT*) controlled by the same p35S, varied, reflecting the expected dissimilar transcript stability and turnover for different sequences. Furthermore, other effects also determined the transcript accumulation because the steady-state transgene transcript levels in transformants containing two copies of the *GUS* transgene were lower than in those carrying only one copy, while in transformants with two copies of the *SPT* or the *GFP* genes, the levels were higher than those with only one of the respective genes. The transcript length and GC content of the coding sequences were unlikely to be primary determinants for the gene-specific threshold for the silencing trigger [69]. Indeed, although the *GFP* and *SPT* transgenes have almost identical transcript lengths and similar GC content, silencing was activated at a different copy number. Silencing of *GUS* expression was already observed in plants with three P35S-*GUS* copies, while *GFP* and *SPT* expression were only reduced in plants carrying five or more 35S-*GFP* and nine or more 35S-*SPT* copies.

In tobacco plants, different 5' leader sequences were modulated by P35S-driven expression [6].

A 31-nucleotide random leader stimulated translation 20- and 100-fold compared to the nine- and four-nucleotide synthetic leaders, respectively. However, the 30-nucleotide satellite tobacco necrosis virus leader and both the 79-nucleotide tobacco mosaic virus (TMV) and the plant chlorophyll *a/b*-binding protein (66-Cab22L) leaders were approximately two- to three-fold and fivefold stronger than this 31-nucleotide random leader, respectively [6]. On the contrary, the 5'-untranslated regions (5' UTRs) of MAS and TMV did not influence the transgene expression variability [66]. Similarly, also four terminators (tMAS, tNOS, tG7, and tOCS) were analyzed for their effect on transgene expression, but none affected levels and variability [66]. That especially certain combinations of different elements can lead to high and stable transgene expression was clearly demonstrated [70]. The combination of the strong β -phaseolin promoter, with the 5' and 3' regulatory sequences of the *arceline 5-I* gene and efficient endoplasmic reticulum-targeting signals boosted the expression of an antibody fragment to 36% of the total soluble protein in *Arabidopsis* seeds. Furthermore, the variation in transgene expression was very low [70].

A significantly increased level of transgene expression can be observed for particular intron-containing transgenes when compared to their respective intronless constructs. This phenomenon, designated "intron-mediated enhancement," has been demonstrated both in monocots and dicots [71–74]. The ability of an intron to enhance transgene expression depends on the sequence and position of the intron within the transgene [72, 75]. Generally, stimulation of mRNA accumulation decreases with increasing distance of the intron from the promoter and its effect is greatly diminished or entirely lost when the intron is placed in the 3' UTR [71, 75]. The mechanism of intron-mediated enhancement is largely unknown, but has been postulated to be due to an increased accumulation of mRNA, enhanced transcript stability, and/or an enhanced translation [71, 72, 74, 76].

The effect of matrix attachment regions (MARs) on transgene expression levels and stability is still unclear and contradictory results have been obtained [66, 77–79]. In some studies, MARs significantly increased the average protein accumulation levels in transgenic plants and reduced the transgene expression variability [80–83]. For instance, the chicken lysozyme

MAR (chiMAR) reduced the expression variability by seven- to eightfold in transgenic potato [83], whereas this chiMAR had no influence on the transgene expression levels and variability in *Arabidopsis* transformants, unless the gene silencing mutants were used as genetic background for transformation [3, 66, 78, 84]. The lack of positive effect of the MARs was caused neither by the transformation method, nor by the plant species used for transformation, because this chiMAR did not influence the *Arabidopsis* transformants obtained after root or floral dip transformation and the tobacco transformants [78]. Moreover, a tobacco-derived MAR sequence had no positive effect on transgene expression in the *Arabidopsis* wild-type Columbia-0 transformants [78].

The Influence of the Integration Position on Transgene Expression

Although the presence of multiple transgene copies, especially in inverted orientation, is the major trigger to induce transgene silencing (see below), also single-copy plants can show variation in transgene expression and undergo transgene silencing [5, 17, 60, 85–88]. Two general explanations for the inactivation of single-copy transgenes have been proposed that are not mutually exclusive: recognition of a transgene as foreign DNA and subsequent TGS and neighborhood of spreading heterochromatin domains. In the first explanation for position effects, transgenes might be recognized as nonself sequences, and especially the difference in GC content of the transgene versus the flanking sequences might induce chromatin changes and epigenetic variation [12, 89]. Indeed, a single copy of the GC-rich *A1* gene from maize became specifically methylated in transgenic petunia (*Petunia hybrida*), while the homolog of gerbera (*Gerbera hybrida*), with a GC content similar to that of petunia, remained unmethylated [89, 90]. In the second explanation, position effects result from heterochromatinization of the transgene integrated close to heterochromatin domains, which is similar to the position effect variegation, extensively studied in fruit fly (*Drosophila melanogaster*). Thus, also in plants, the chromosomal position of a transgene is expected to cause variability in transgene expression [9, 91]. In petunia, single-copy transgenes were strongly silenced

when they were integrated into a highly repetitive region [5]. Likewise, two single-copy transformants were identified with unstable transgene expression that were integrated into intercalary and paracentromeric locations [85], while the two stably expressing loci were integrated into the vicinity of telomeres and flanked at least on one side by plant DNA with AT-rich regions. However, the number of examined cases is limited and T-DNA insertion into heterochromatic regions has been demonstrated not to be necessarily associated with transgene silencing [16].

To determine whether the integration position had a major effect on transgene expression variability, a large transformant population had been screened for single-copy T-DNA transformants and the transgene accumulation levels were compared [17]. Twenty single-copy transformants were selected based on antibiotic resistance marker expression and in all of them the T-DNA integration was characterized at the sequence level in the *Arabidopsis* genome (Fig. 1a). In 18 of the 20 transformants, the p35S-driven *GUS* expression was high and stable in two subsequent generations. Furthermore, the *GUS* activity levels in these 18 different transformants could be considered as similar, because the intra-transformant variability was as high as the inter-transformant variability (Fig. 1b). Integration into an intergenic or genic region and into an exon or intron did not result in differential transgene expression (Fig. 1) and, additionally, integration into a gene in sense or antisense orientation had no influence on the transgene expression, indicating that overlapping transcription from the endogene and transgene did not induce transgene silencing in *Arabidopsis* [17].

Another hypothesis was that transgene expression might be reduced by the absence of MAR sequences near the T-DNA integration site or by the presence of neighboring highly methylated sequences [92, 93]. However, no general correlation with these factors was observed [17]. There has also been some debate on the influence of vector backbone sequences, cointegrated into the plant genome on transgene expression. In a first report based on the observation that two unstably expressed loci harbored binary vector sequences directly contiguous to the right T-DNA border, whereas two stably expressed loci contained no vector backbone sequences, vector sequences were postulated to trigger TGS [85]. Later, several studies

showed that the integration of vector sequences, even with a GC content strongly diverged from that of the plant genome, had no negative influence on transgene expression [17, 94, 95].

In general, one can conclude that in *Arabidopsis*, the integration position of single-copy T-DNAs does not strongly influence transgene expression [17, 62, 94], but one should take care not to extrapolate this finding too far. The *Arabidopsis* genome is small (125 Mb) and has a low amount of heterochromatic DNA, while the genome of *Drosophila*, in which position effect variegation is often observed, is comparable in size, but consists for one third of centric heterochromatin [96]. Also, one should be aware that all the above-described observations were based on transformants that were selected on the expression of one selectable marker on the same T-DNA, implying that silenced heterochromatic T-DNA insertions were not recovered. Indeed, an increased number of T-DNA integrations into heterochromatic regions were found without selection in *Arabidopsis* [20, 21]. Apparently, the kanamycin selection had failed to identify approximately 30% of all integration events, because the expression of the selectable marker was absent or very low [21]. Analysis revealed that not PTGS, but TGS, caused this discrepancy in transformation efficiency. Furthermore, the integration sites of the lines with silenced transgenes mapped to heterochromatic regions, including telomeres, centromeres, and rDNA repeats [20, 21], which are regions that are significantly underrepresented in T-DNA integration studies with transformants identified by selection [21].

Nevertheless, the holy grail of plant biotechnology is to integrate the transgenes into the plant genome via homologous recombination. Indeed, in contrast to yeast and the moss *Physcomitrella patens*, the frequency of transgene integration via homologous integration is extremely low. When compared to random integration, targeted integration of a transgene into the homologous sequence of the plant genome occurs in the range from 0.01% to 0.1% [97, 98]. Several approaches, with variable success, have already been applied to detect gene targeting events among the random integration events: (1) application of gene-specific selection or screening for the target genes; (2) use of positive-negative selection and, hence, reduction of the transformants with randomly integrated transgenes; and (3) overexpression of

genes involved in homologous recombination in yeast, such as the *RAD54* gene [99–101]. Additionally, induction of double-stranded breaks in the genome has been reported to result in an increased targeted integration frequency [102, 103]. Nowadays, these double-stranded breaks are produced by zinc finger nucleases (ZFNs) [104–106]. These ZFNs are synthetic restriction enzymes that can be specifically designed to cleave virtually any long stretch of double-stranded DNA sequence [107]. A defective *GUS:NPTII* reporter gene, carrying a recognition site for ZFN and integrated at various chromosomal sites in ten different tobacco lines, was restored via homologous recombination in 10% of the transformed protoplasts, regardless of the chromosomal position of the reporter gene [105]. Approximately 20% of the *GUS:NPTII* reporter gene was repaired solely with homologous recombination, but the other events still contained additional DNA insertions and deletions [105]. Moreover, ZFNs are probably a promising tool to efficiently achieve targeted integration into plants. Once this procedure is optimized, plants with high, stable, and predictable accumulation levels of important heterologous proteins might be obtained with an increased frequency.

Correlation Between T-DNA Locus Structure and Homology-Based Silencing

Based on the gene dosage rationale, screening for transgenic plants with multiple transgene copies might be predicted to result in plants with high transgene expression. A direct correlation between transgene copy number and expression level has indeed been reported [51, 62]. However, also the inverse correlation is observed with p35S-driven transgenes inserted as convergently transcribed inverted repeats: the reproduction of the expression of the inverted repeat transgene copies is much lower than that of a single-copy transgene [51, 108].

Increased Transgene Dosage Can Result in PTGS

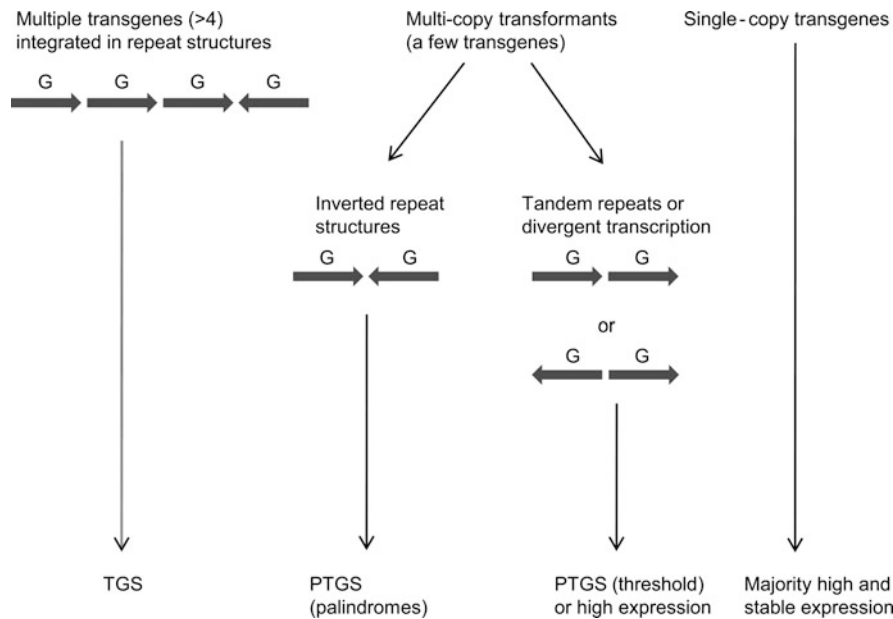
Below a certain number of identical p35S-driven transgenes, gene copy number and expression correlated positively [62]. Transgene expression seemed stable and high over all generations analyzed and a comparable expression level was obtained for all independent lines harboring the same copy number of a certain

transgene. However, once above a certain copy number, gene silencing occurred, implying that silencing was triggered by threshold concentrations of either transgene transcript or another product of transgene expression (Fig. 2) [62]. In many reports, the correlation between high transgene copy number and transgene expression was negative: the higher the copy number, the lower the expression level per gene copy [59, 61, 109]. Not the copy number per se, but especially the arrangement of the transgene copies in one genetic locus, seemed responsible for the low transgene expression. Indeed, as described above, multiple T-DNAs at one genetic locus are frequently integrated as a direct or an inverted repeat (Fig. 2) [1, 34, 35].

Several other examples show a correlation between transgene silencing and the presence of tandem repeats, which might be linked to exceeding the mRNA threshold level above which PTGS is initiated (Fig. 2) [59, 110, 111]. The same is true for transgene loci that are not silenced under hemizygous, but become silenced under homozygous conditions, as, for instance, in tobacco and *Arabidopsis* [60, 87, 112, 113].

Convergent Transcription from Inverted T-DNA Repeat Structures Is a Strong Trigger for PTGS

Invariably, convergently transcribed 35S-driven transgenes from invertedly repeated T-DNAs are posttranscriptionally silenced to a level that is only 1% or less of the single-copy transgene (Fig. 2) [4, 51, 59, 61, 108, 113–118]. Additionally, invertedly repeated transgenes can very efficiently silence *in trans* homologous sequences located elsewhere in the genome [51, 108, 117, 119]. This direct correlation between invertedly repeated transgenes and transgene silencing has been demonstrated experimentally by the deletion of one of the transgenes from the inverted repeat that released transgene silencing and a 100-fold increase in expression [108]. Indeed, in two parental lines, harboring two invertedly repeated 35S-driven *GUS* genes, the *GUS* activity was low and both the coding sequences and the center of the inverted repeat were densely methylated. Homologous transgenes at other chromosomal positions were silenced *in trans* by the inverted repeat silencer locus [108, 118]. Removal of one of the *GUS* copies from this silencer locus by the CRE recombinase resulted in all cases in constitutively



Transgene Expression in Plants, Control of. Figure 2

Relation between locus structure and the expression of a p35S-driven transgene. Upon transformation, both single-copy and multicopy transformants are generated. Both single-copy plants and plants with multiple T-DNAs arranged in tandem repeat or in repeats giving rise to transgene divergent transcription generally display high and stable transgene expression. Plants carrying convergently transcribed transgenes show low transgene expression as a result of PTGS, whereas plants harboring multiple transgenes integrated into a concatemeric array frequently result in TGS.

Abbreviations: G gene, TGS transcriptional gene silencing, PTGS posttranscriptional gene silencing

high *GUS* expression, a significant decrease in methylation and lack of *in trans* silencing capacity [108, 118]. Strikingly, the spacer region in-between the two invertedly repeated transgenes seemed to determine the efficiency of transgene silencing. The presence of an 826-bp non-repetitive spacer region in-between the invertedly repeated transgenes strongly decreased the degree and the stability of silencing [108], but in another system, inverted repeats interrupted by a non-palindromic spacer region varying from 500 bp to 1,022 bp still gave efficient silencing [120].

Multiple T-DNA Copies at One Locus Tend to Become Transcriptionally Silenced After Several Generations

Multiple T-DNA repeats not only induce PTGS, but also TGS. Induction of TGS occurs especially when a concatemeric array or inverted repeats of transgenes are formed and when different genes are regulated by

the same promoter [119, 121–124]. The 271 locus, consisting of multiple T-DNAs with an antisense nitrite reductase driven by the 35S promoter, is transcriptionally silenced and is also able to very efficiently *in trans* silence another single-copy p35S-driven transgene [119]. Transgene expression is seemingly less stable over sexually than over vegetatively propagated generations [125]. Indeed, expression of the sulfur-rich sunflower (*Helianthus annuus*) seed albumin (SSA) and the phosphinothricin (*BAR*) genes, all driven by p35S, progressively decreased in transgenic clover (*Trifolium subterraneum*) [122]. The expression of both genes was stable until the T3 generation, but in the T7 generation, all plants were completely susceptible to the herbicide and the mean level of the SSA protein was much lower than those observed in the T3 generation. This progressive decrease in expression correlated with the reduced transcription level of both genes and strong CpG methylation in the promoter [122]. Not only the strong 35S promoter, but also the weak pNOS can

trigger TGS. The H2 locus, a locus harboring six copies of pNOS, all transcriptionally silenced loci, strongly methylated the promoter regions [121]. Also, transcription of an inverted repeat of pNOS could trigger TGS and methylation of pNOS *in trans*, and this progressively decreased through generations [124].

How to Prevent Transgene Silencing?

To prevent transgene silencing, single-copy transgene inserts should be screened (see below) and inclusion of identical sequences in different transgene constructs should be avoided [126]. Even a homology of 90 bp in the promoter region can be sufficient to trigger TGS [119]. Additionally, the use of multiple identical 3' end regions can result in low transgene expression levels. A homology of 239 nucleotides in the 3' UTR is sufficient to be recognized efficiently by the homology-based RNA degradation machinery [117] and also a homology of approximately 204 nucleotides in the 3' UTR is enough for a single-copy transgene to activate the *in trans* silencing of another transgene with the same 3' UTR [113]. Also designing transgenes without any inverted repeat is important to guarantee high and stable expression. Indeed, upon transcription of these inverted repeats, hairpin structures and double-stranded RNA are formed initiating transgene silencing [127]. Almost complete cosuppression of an endogenous gene in tomato (*Solanum esculentum*) was obtained after transformation with a homologous transgene construct that harbored two upstream inverted repeats in the 5' UTR.

Control of Gene Expression by the Generation of Single-Copy Transformants

As mentioned above, transformants obtained after *Agrobacterium*-mediated gene transfer often contain multiple T-DNA copies in direct or inverted orientations, and the frequency of single-copy transformants is rather low [1, 17]. Because screening for single-copy T-DNA transformants strongly enriches for transformants with stable and high transgene expression [17], single-copy transformants need to be identified in a large pool of plants. Currently, conventional, but time-consuming and labor-intensive methods, such as DNA gel blot analysis and T-DNA fingerprinting are used [128]. Recently, transformation technologies and vectors have been developed to increase the frequency of

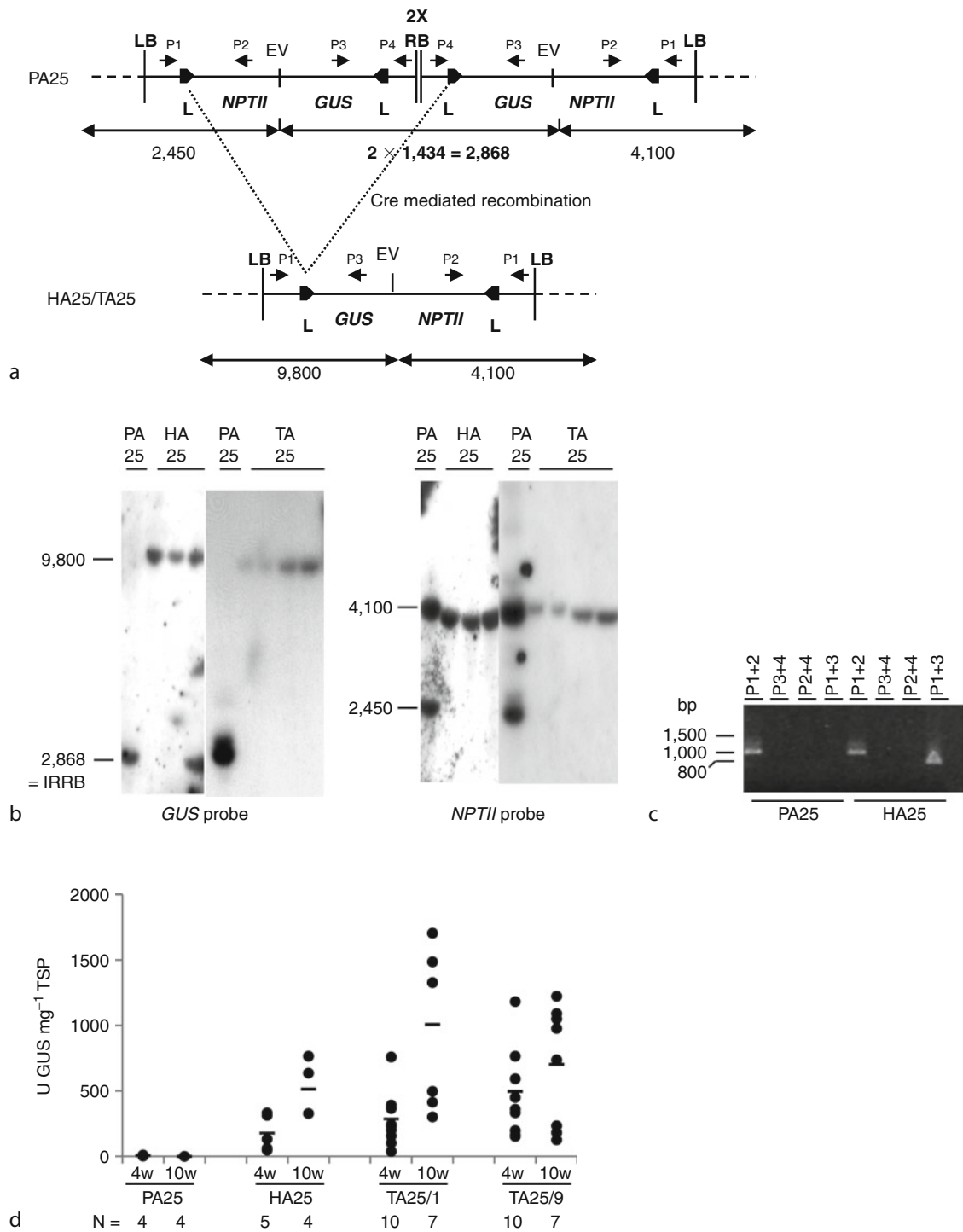
single-copy transformants during or after transformation with site-specific recombination.

Generation of Single-Copy Transformants by Resolution of Complex Integration Loci by Site-Specific Recombination

A T-DNA construct, harboring site-specific *loxP* recombination sites, was designed to reduce complex T-DNA loci into a single T-DNA insert (Fig. 3a) [129]. This T-DNA harbored two invertedly oriented *loxP* recombination sites inside and immediately adjacent to the left and right T-DNA border ends [129]. Recombination between the outermost *loxP* sequences in direct orientation should resolve multiple copies into a single T-DNA copy, regardless of the number and orientation of the *loxP*-derived T-DNA copies inserted at one locus.

In a first approach, seven *Arabidopsis* transformants with multiple T-DNA inserts on a single locus were crossed with a homozygous *CRE*-expressing line [129]. In three hybrids, the complex T-DNA locus was reduced efficiently to a single-copy locus and in two of them, the resolution of the inverted repeat locus was accompanied by an at least tenfold-enhanced and stable transgene expression (Fig. 3a–d). In the progeny plants, only the simplified T-DNA locus was detected upon segregation of the *CRE* recombinase gene, proving that excision took place in the progenitor cells of the gametes (Fig. 3b and c) [129]. Unfortunately, in four of the seven transformants, the complex T-DNA locus could not be resolved by *CRE*-mediated transformation to a single T-DNA copy, although some rearrangements occurred as demonstrated on DNA gel blots. Strikingly, these complex T-DNA loci that were not resolved, had variable expression levels in different progeny plants with the same complex locus, implying some epigenetic imprints imposed by the interaction with the *CRE* recombinase. Possible reasons for the lack of resolution in these transformants might be deletion of the most extreme *loxP* sequences, too low *CRE* activity, or inaccessibility of the *loxP* sequences due to heterochromatinization of the complex locus [129]. Indeed, a correlation between the efficiency of *CRE*-mediated recombination and the *CRE* mRNA levels had been demonstrated [130].

In a second approach, the T-DNA vector with oppositely oriented *loxP* sequences was transformed into



Transgene Expression in Plants, Control of. Figure 3 (Continued)

CRE-expressing *Arabidopsis* plants [88] and 55% of the primary transformants were single-copy T-DNA plants versus only 15% in control plants. Most of the single-copy transformants in *CRE*-expressing background (70%) displayed a continuous somatic inversion of the DNA fragment between the two inverted *loxP* sequences. To avoid this phenomenon, a new T-DNA vector, harboring only one *loxP* sequence adjacent to the LB or RB, was transformed in a *CRE*-expressing plant and up to 70% of the transformants were single copy. As resolution to single-copy plants is only efficient when the outermost *loxP* sites are in direct orientation, inverted T-DNA repeats might be mostly internal to multiple T-DNA copy arrays and they rarely occur at the ends after floral dip transformation in *Arabidopsis* [88]. Most single-copy transformants, obtained with both T-DNA vectors, displayed high and stable transgene expression. Strikingly however, the transgene expression in the majority of multicopy *CRE* transformants was stable and uniform, suggesting that the *CRE* recombinase prevented also the generation of inverted T-DNA repeats or modified the chromatin structure of the locus, so that it became less sensitive for gene silencing [88]. In both strategies, the *CRE* expression cassette is still present after

resolution of the complex T-DNA locus, which can be a disadvantage, because, except for *Arabidopsis*, high *CRE* expression can result in severe growth phenotypes and reduced fertility in some plant species [131]. To avoid this drawback, *loxP*-containing T-DNAs could be transformed into hemizygous *CRE*-expressing plants, with 25% of the transformed progenies without *CRE* expression cassette as a consequence [88]. Also backcrossing of a homozygous *CRE*-expressing line with a wild-type plant would result in transformants without the *CRE*-containing T-DNA. Alternatively, the *CRE* recombinase protein could also be transiently introduced into the plant cell nucleus. Vergunst et al. [132] developed a VirB/D4-dependent translocation system in which the *CRE* recombinase was fused with the VIRE2 or VIRF proteins of *A. tumefaciens*. Together with the VIR proteins, the *CRE* recombinase was transferred to the plant cell and site-specific recombination took place. This recombination did not require any transferred DNA [132], but the question remains whether this *CRE* recombinase activity would be high enough to mediate complex T-DNA loci resolution and to transfer the *CRE* recombinase very efficiently.

Transgene Expression in Plants, Control of. Figure 3

Single-copy T-DNA transformants with high and stable transgene expression as a result of *CRE/loxP* recombination-mediated resolution (Adapted from [129]). (a) Schematic representation of the *CRE/loxP*-mediated resolution of a complex locus to a single T-DNA (not on scale). Parental (P) plant PA25 contained two T-DNAs in inverted orientation. After the plant DNA was digested with the *EcoRV* (EV) enzyme, a fragment of 2,868 bp after hybridization with the *GUS* probe is indicative for the inverted T-DNA repeat about the RB, while two T-DNA/plant junctions of 2,450 bp and 4,100 bp were observed after hybridization with the *NPTII* probe, revealing the left T-DNA/plant DNA junctions. The hybrid (H) HA25 was obtained after crossing the parental plant with a *CRE*-expressing plant. *CRE/loxP*-mediated resolution between the two outermost *loxP* sites resulted in a new T-DNA harboring two LB regions in HA25 and the F2 progeny plants TA25. The primers P1, P2, P3, and P4 are indicated, showing homology with the LB region, the *GUS*-coding sequence, the *NPTII* coding sequence, and the RB region, respectively. (b) DNA gel blot analysis on *EcoRV*-digested DNA of PA25, HA25, and TA25. After *CRE/loxP*-mediated resolution, only one original LB-T-DNA/plant junction fragment was detected after hybridization with the *NPTII* probe fragment. Additionally, as expected, one new T-DNA/plant junction was observed after hybridization with the *GUS* probe. (c) Polymerase chain reaction (PCR) analysis on PA25 and HA25 transformants to identify the newly formed T-DNA configurations. Primer combination P1+2 amplified an LB/*NPTII* junction of 1,142 bp; primer combination P1+3 amplified an LB/*GUS* junction of 914 bp (arrow 2). With primer P4, no PCR fragment was obtained in PA25, because no inverted repeats could be amplified by PCR. Upon *CRE/loxP*-mediated resolution, the RB inverted palindrome was deleted, and only PCR products with the LB region of primer P1 were obtained. (d) *GUS* activity analysis in leaves of 4- and 10-week-old PA25, HA25, and TA25 seedlings. The measurements and the mean are indicated by dots and line, respectively. *GUS* activity levels are given as units GUS mg^{-1} of total soluble protein. The number (*N*) of the analyzed seedlings is indicated

Simplification of complex T-DNA loci was also observed by Verweire et al. [133]. To ultimately develop a T-DNA vector that generates homozygous marker-free transgenic *Arabidopsis* plants without the need of additional transformation rounds or crosses, a germline-specific auto-excision vector was designed that harbored both the selectable marker and the *CRE* expression cassette between two *loxP* sequences in direct repeat. Additionally, the *CRE* recombinase was driven by the germline-specific 2.2-kb promoter fragment of the *SOLO DANCERS* gene, which is active in both the male and female meiocytes of *Arabidopsis*. In this manner, the transgenic plants become genetically programmed so that the marker gene is lost after the initial selection of the primary transformants. Surprisingly, not only marker-free homozygous progenies were obtained very efficiently, but also the locus structure in these progeny plants was simplified [133].

A similar approach to resolve complex transgene loci had been developed previously for the generation of single-copy transgenic wheat plants after direct gene transfer. In this case, the selectable marker was removed simultaneously with the resolution of the complex locus [134]. The transformation vector consisted of a transgene flanked by two *lox511* sites, a mutant variant of the wild-type *loxP* sequence, in inverted orientation. Additionally, the selectable marker gene was flanked by two wild-type *loxP* recombination sites in the same orientation. Two correct and independent recombination reactions could take place, because *lox511* and *loxP* did not recombine, with transformants with a marker-free, single-copy insertion of the transgene in the plant genome as a consequence. Crossing of four wheat transformants, harboring a multicopy locus, with a *CRE*-expressing wheat transformant, resulted in 62/72 F₂ plants with a single-copy, marker-free, transgene locus [134]. Because the resolved locus still contained two *lox511* sites in inverted orientation, the *CRE* recombinase could still mediate inversion of the transgene between these two sites. Therefore, all F₁ plants were chimeric for the inversion and harbored both transgene orientations.

To avoid crossing with a *CRE*-expressing plant, the transgene, flanked by the oppositely oriented *loxP* sites in maize cells, was cobombarded with a *CRE*-expressing construct [135]. This cotransformation

resulted in 85% primary transformants harboring one or two copies of the introduced gene, of which 38% harbored a single copy. Of these single-copy plants, 60% lacked also the recombinase gene, because during cobombardment a molar ratio of 3:1 transgene to *CRE* construct was used. In this manner, an overall efficiency of 23% of plants with only one copy of the transgene construct was obtained [135]. This cotransformation strategy requires that the selectable marker is retained in the primary transformants. However, a strategy with two different recombination systems and two different selectable markers could result in the integration of one transgene without the incorporation of additional unneeded DNA in the transgenic plant [135].

Besides the use of the *CRE/loxP* recombination system to obtain efficiently single-copy transformants, other strategies have been proposed and evaluated. A site-specific recombination strategy was described in which two sets of directly repeated FLP recognition target (FRT) sites [136] flanked a “to-be-removed” (TBR) region. The FRT sites flanking one TBR are inverted in relation to the sites flanking the other TBR. Each TBR can be excised independently, because only recombination between directly repeated FRT sites can cause excision. Furthermore, one site-specific recombinase reaction could simultaneously induce double excisions, resulting in the resolution of complex T-DNA loci. Indeed, the progenies of 70% of the *Arabidopsis* transformants harbored only one or two copies of the transgene, while 40% contained only one T-DNA copy. This frequency of transformants with a simple T-DNA integration pattern is significantly higher than the 5–20% single-copy transformants normally obtained after floral dip transformation [1, 17]. Most interestingly, the reduced copy number was also strongly associated with increased expression [137].

Generation of Single-Copy Transformants by Site-Specific or Site-Directed Integration

Besides the use of site-specific recombination to resolve complex transgene loci into simple loci before, during or after integration, this site-specific recombination system can also be applied to produce plants with a simple integration pattern through site-specific integration [86, 138–141]. In addition, with this technology, transgenes can be targeted to a specific preselected chromosomal site

to eliminate variation in gene expression. The strategy to obtain single-copy transformants via site-specific integration consists of two transformation rounds. In the first round, a target line is generated, harboring a target *lox* site between the promoter and the *CRE* gene. In the second transformation, the vector carrying the gene of interest and two *lox* sites is introduced into the target line. When the gene of interest is integrated into the *lox* site of the target line, the expression of the recombinase gene is stopped. Site-specific gene integration in plants has been described with various DNA delivery methods, such as polyethylene glycol-mediated tobacco protoplast transformation [86], *Agrobacterium*-mediated transformation of *Arabidopsis* [138], and biolistic-mediated transformation of rice [139, 141]. After *Agrobacterium*-mediated transformation, the frequency of *CRE*-mediated gene integration was very low (approximately 2%) compared to the DNA delivery via direct gene transfer, probably because the transferred DNA during *Agrobacterium* transformation is single stranded, which is not a substrate for the *CRE* recombinase [139, 140].

By combining site-specific gene integration with gene expression analysis, the absolute level of transgene expression could be shown to vary up to tenfold, depending on the target site in the tobacco genome, indicating that the chromosome position can affect the level of transgene expression [86]. Additionally, despite the identical integration pattern, half of the transgenic tobacco lines showed the expected expression of the transgene and the other half very low *GUS* expression. This low *GUS* expression was due to transgene silencing, because it was correlated with DNA methylation and low transcript levels. Furthermore, this DNA methylation was specific for the newly introduced DNA sequences. Also in independent experiments, regenerants could be divided also upon site-directed integration in two categories: the single-copy lines that contained one site-specific insert without additional sequence integration and multicopy lines in which, besides the site-specific integration, additional transgene copies were integrated into the plant genome [140, 141]. The transgene expression of the single-copy rice transformants was high and stable and the variation was lower than that of the multicopy lines. In approximately half of the multicopy lines, expression of the site-specific integrated transgene

could be reactivated and stabilized when the illegitimate integrated inserts were segregated away. In the remaining half, the expression could not be restored, because the random integrations were genetically linked to the site-specific integration event [141].

A site-directed integration system for *Agrobacterium*-mediated transformation of tobacco was developed in which a single transgene copy could be integrated precisely into a predefined target locus by recombinase-mediated cassette exchange (RMCE) [142]. In the RMCE-based site-directed integration strategy that uses the R-RS system from *Zygosaccharomyces rouxii*, a single-copy of the target cassette, surrounded by two oppositely oriented RS sites, is randomly integrated into the plant genome. The exchange cassette contains the selectable marker gene and the gene of interest between two opposite RS sites, and the recombinase gene, the selectable marker gene, and the gene of interest between directly oriented RS sites. This third RS site excluded random integration events. The recombinase will catalyze a double crossover between the two RS sites, replacing the target cassette by the exchange cassette and removing the recombinase gene. Expression analysis revealed that the obtained site-specific recombined plants from the same target line had approximately the same transgene expression level and less transgene expression variability than the random-integration transgenic plants and no transgene silencing [142]. Strikingly, transgenes in the same direction at the same target locus had the same level of activity, in contrast to transgenes in different directions, indicating that the surrounding genome DNA sequence outside the target locus might affect the activity of the gene. Also after direct gene transfer of soybean (*Glycine max*), site-specific integration via RMCE occurred efficiently and stable transgene expression was stable [143].

Another approach to obtain single-copy transgenic lines is the “Agrolistic” method, in which the advantages of *Agrobacterium*-mediated transformation and biolistics were combined [144, 145]. The *VIRD* genes were cobombarded with the T-DNA borders, flanking the introduced transgene. Approximately, 20% of the transformed tobacco calli contained only the T-DNA with correctly processed border sequences and no integrated vector backbone sequences [144], whereas

10–35% of transgenic maize calli contained one to two transgene copies. The addition of *VIRE2* genes even doubled this transformation efficiency [145]. Finally, the application of niacinamide, a product that reduces recombination of extrachromosomal molecules, resulted in the generation of single-copy transformants in wheat with an efficiency of 8% [146].

Use of Gene Silencing Mutants to Overcome Transgene Expression Variability

Variation in transgene expression in a population of plants is frequently due to PTGS of the transgenes. Several genes are involved in PTGS and mutant screens revealed that the RNA-dependent RNA polymerase (RDRP6, SGS2) is essential for PTGS to occur [147, 148]. Therefore, a logical approach to overcome PTGS in a population of transgenic plants was to generate transformants in an *rdp6* mutant background, impaired in PTGS [3, 84]. Introduction of p35S-*GUS* or p35S-*GFP* transgenes into the *Arabidopsis* *sgs2* and *sgs3* mutants (two different alleles of *RDR6*) resulted in stable and high expression of these transgenes in 100% of the transformants analyzed [3], while, similarly, the outcome of introduction of a cyclin-dependent kinase inhibitor 6 gene under the control of p35S was a clear overexpression phenotype, typical for high expression of the kinase inhibitor, in all transformants. Also the transgene expression in transformants with convergent transcription of invertedly repeated T-DNAs was high. Interestingly, immediate transgene silencing in the F1 progeny plants occurred by backcrossing of these high-expressing transformants containing an inverted repeat T-DNA locus with a wild-type nontransformed plant. This observation clearly demonstrates that RDRP6 is required for the generation of double-stranded RNA, also from inverted repeat loci, and that not the read-through transcripts from the convergently transcribed genes generate the double-stranded RNA, as is widely believed. Similar results were obtained when the transgenes were driven by the pCAS promoter, a strong constitutive promoter derived from the cassava vein mosaic virus [149]. On the contrary, expression variability was not significantly reduced in the plants transformed with *sgs2* and *sgs3* when the transgenes were under control of the hybrid pOMA1 promoter, a hybrid promoter consisting of

parts of the pMAS and pOCS, which might be too weak to trigger PTGS [3], in contrast to the strong p35S and pCAS. When the T-DNA constructs were flanked by MARs of the chicken lysozyme gene, the GUS activity was boosted fivefold and 12-fold in *sgs2* and *sgs3* background, respectively, in contrast to the wild-type background, making this PTGS-MAR expression system very attractive for the production of heterologous proteins in plants [3, 84].

Future Directions

The ultimate aim is to generate transgenic plants with a predictable, well-controlled, and stable transgene expression over many generations. As described above, to achieve this goal, many parameters should be taken into account. Most importantly, the transgene construct should be optimally designed: a suitable promoter should be selected that drives transgene expression in the desired way, multiple homologous sequences should be avoided in the transgene construct, and foreign unnecessary DNA should be as much as possible prevented to be integrated into the plant genome. Until now, the transgene construct was integrated via illegitimate recombination, but recent reports on site-specific integration let one assume that targeted integration via ZFNs will soon become feasible routinely [105]. At the same time, this method will enrich for single-copy transformants ensuring stable transgene expression. Importantly, the majority of the above-described observations and conclusions are all based on genes driven by the strong constitutive p35S. Therefore, it is perfectly conceivable that all these observations are typical for this promoter, but cannot be extrapolated to other promoters. In the future, the use of strong plant promoters instead of the 35S promoter in transgene constructs might be a big improvement to obtain controllable and stable high transgene expression. Indeed, in general, plant promoter–gene fusions show limited quantitative and qualitative variations in transgene expression in different transformants. Analysis of the expression profile of the *ARCELIN-5* regulatory sequences in *Arabidopsis* seeds revealed that, surprisingly, the *Arabidopsis* seeds accumulated ARCELIN-5, an abundant protein found in common bean (*Phaseolus vulgaris*), to 15% of the total soluble protein content [150]. All transformants

had low plant-to-plant variation and even in plants harboring a complex T-DNA integration pattern, with invertedly and directly repeated T-DNAs, the transgene expression was high and stable [150]. Similar results were found for genes driven by the β -PHASEOLIN promoter of common bean [70].

However, to really control transgene expression in plants, the underlying triggers that initiate PTGS and TGS have to be understood in much more detail. Do all plant promoters give less variation in transgene expression compared to viral and bacterial promoters? Do plant promoters also induce silencing and with which frequency? Intriguingly, why do certain loci with multiple p35S-driven transgene copies provoke silencing and why others not? Silencing is often triggered when gene expression rises above a certain threshold, but what is the nature of this threshold and how is the threshold measured? Can transcript levels be increased by the insertion of particular stabilizing elements? To answer all these questions, an integrated approach will have to be followed, including the use of well-designed transgene constructs differing only in a single transgene element in combination with different mutants. Molecular characterization of the transgene locus, the transcript steady state and turnover levels, and the translation efficiency should allow to unravel the network of gene expression control and identify the most important master elements. The ultimate goal is to be able to construct transgenes with a predefined expression pattern in a reliable way.

Acknowledgments

The authors thank Martine De Cock for help with the manuscript. This work was supported in part by grants from the European Union BIOTECH program (QLRT-2000-00078) with additional cofinancing from the Flemish Community, the “Bijzonder Onderzoeksfonds” of Ghent University (BOF 01111400), and the Research Foundation-Flanders (no. G.021106).

Bibliography

- De Buck S, Podevin N, Nolf J, Jacobs A, Depicker A (2009) The T-DNA integration pattern in *Arabidopsis* transformants is highly determined by the transformed target cell. *Plant J* 60:134–145
- Windels P, De Buck S, Depicker A (2008) *Agrobacterium tumefaciens*-mediated transformation: patterns of T-DNA integration into the host genome. In: Tzfira T, Citovsky V (eds) *Agrobacterium: from biology to biotechnology*. Springer, New York, pp 441–481
- Butaye KMJ, Goderis IJWM, Wouters PFJ, Pues JM-TG, Delauré SL, Broekaert WF, Depicker A, Cammue BPA, De Bolle MFC (2004) Stable high-level transgene expression in *Arabidopsis thaliana* using gene silencing mutants and matrix attachment regions. *Plant J* 39:440–449
- Muskens MWM, Vissers APA, Mol JNM, Kooter JM (2000) Role of inverted DNA repeats in transcriptional and post-transcriptional gene silencing. *Plant Mol Biol* 43:243–260
- Pröls F, Meyer P (1992) The methylation pattern of chromosomal integration regions influence gene activity of transferred DNA in *Petunia hybrida*. *Plant J* 2:465–475
- De Loose M, Danthinne X, Van Bockstaele E, Van Montagu M, Depicker A (1995) Different 5' leader sequences modulate β -glucuronidase accumulation levels in transgenic *Nicotiana tabacum* plants. *Euphytica* 85:209–216
- Napoli C, Lemieux C, Jorgensen R (1990) Introduction of a chimeric chalcone synthase gene into petunia results in reversible co-suppression of homologous genes *in trans*. *Plant Cell* 2:279–289
- van der Krol AR, Mur LA, Beld M, Mol JNM, Stuitje AR (1990) Flavonoid genes in petunia: addition of a limited number of gene copies may lead to a suppression of gene expression. *Plant Cell* 2:291–299
- Depicker A, Van Montagu M (1997) Post-transcriptional gene silencing in plants. *Curr Opin Cell Biol* 9:373–382
- Frizzi A, Huang S (2010) Tapping RNA silencing pathways for plant biotechnology. *Plant Biotechnol J* 8:655–677
- Vaucheret H, Fagard M (2001) Transcriptional gene silencing in plants: targets, inducers and regulators. *Trends Genet* 17:29–35
- Matzke AJM, Matzke MA (1998) Position effects and epigenetic silencing of plant transgenes. *Curr Opin Plant Biol* 1:142–148
- Tzfira T, Citovsky V (2006) *Agrobacterium*-mediated genetic transformation of plants: biology and biotechnology. *Curr Opin Biotechnol* 17:147–154
- Banta LM, Montenegro M (2008) *Agrobacterium* and plant biotechnology. In: Tzfira T, Citovsky V (eds) *Agrobacterium: from biology to biotechnology*. Springer, New York, pp 73–148
- Tzfira T, Li J, Lacroix B, Citovsky V (2004) *Agrobacterium* T-DNA integration: molecules and models. *Trends Genet* 20:375–383
- Forsbach A, Schubert D, Lechtenberg B, Gils M, Schmidt R (2003) A comprehensive characterization of single-copy T-DNA insertions in the *Arabidopsis thaliana* genome. *Plant Mol Biol* 52:161–176
- De Buck S, Windels P, De Loose M, Depicker A (2004) Single-copy T-DNAs integrated at different positions in the *Arabidopsis* genome display uniform and comparable β -glucuronidase accumulation levels. *Cell Mol Life Sci* 61:2632–2645

18. Koncz C, Martini N, Mayerhofer R, Koncz-Kalman Z, Körber H, Redei GP, Schell J (1989) High-frequency T-DNA-mediated gene tagging in plants. *Proc Natl Acad Sci USA* 86:8467–8471
19. Herman L, Jacobs A, Van Montagu M, Depicker A (1990) Plant chromosome/marker gene fusion assay for study of normal and truncated T-DNA integration events. *Mol Gen Genet* 224:248–256
20. Kim S-I, Veena GSB (2007) Genome-wide analysis of *Agrobacterium* T-DNA integration sites in the *Arabidopsis* genome generated under non-selective conditions. *Plant J* 51:779–791
21. Francis KE, Spiker S (2005) Identification of *Arabidopsis thaliana* transformants without selection reveals a high occurrence of silenced T-DNA integrations. *Plant J* 41:464–477
22. Windels P, De Buck S, Van Bockstaele E, De Loose M, Depicker A (2003) T-DNA integration in *Arabidopsis* chromosomes: presence and origin of filler DNA sequences. *Plant Physiol* 133:2061–2068
23. Wenck A, Czakó M, Kanevski I, Márton L (1997) Frequent collinear long transfer of DNA inclusive of the whole binary vector during *Agrobacterium*-mediated transformation. *Plant Mol Biol* 34:913–922
24. Kononov ME, Bassuner B, Gelvin SB (1997) Integration of T-DNA binary vector 'backbone' sequences into the tobacco genome: evidence for multiple complex patterns of integration. *Plant J* 11:945–957
25. De Buck S, De Wilde C, Van Montagu M, Depicker A (2000) T-DNA vector backbone sequences are frequently integrated into the genome of transgenic plants obtained by *Agrobacterium*-mediated transformation. *Mol Breed* 6:459–468
26. Podevin N, De Buck S, De Wilde C, Depicker A (2006) Insights into recognition of the T-DNA border repeats as termination sites for T-strand synthesis by *Agrobacterium tumefaciens*. *Transgenic Res* 15:557–571
27. van der Graaff E, den Dulk-Ras A, Hooykaas PJJ (1996) Deviating T-DNA transfer from *Agrobacterium tumefaciens* to plants. *Plant Mol Biol* 31:677–681
28. Barcelo P, Lazzeri PA (1998) Direct gene transfer: chemical, electrical and physical methods. In: Lindsey K (ed) *Transgenic plant research*. Harwood Academic Publishers, Amsterdam, pp 35–55
29. Kohli A, Leech M, Vain P, Laurie DA, Christou P (1998) Transgene organization in rice engineered through direct DNA transfer supports a two-phase integration mechanism mediated by the establishment of integration hot spots. *Proc Natl Acad Sci USA* 95:7203–7208
30. Zhao ZY, Gu W, Cai T, Tagliani LA, Hondred D, Bond D, Krell S, Rudert ML, Bruce WB, Pierce DA (1998) Molecular analysis of T0 plants transformed by *Agrobacterium* and comparison of *Agrobacterium*-mediated transformation with bombardment transformation in maize. *Maize Genet Coop Newsletter* 72:34–37
31. Dai S, Zheng P, Marmey P, Zhang S, Tian W, Chen S, Beachy RN, Fauquet C (2001) Comparative analysis of transgenic rice plants obtained by *Agrobacterium*-mediated transformation and particle bombardment. *Mol Breed* 7:25–33
32. Shou H, Frame BR, Whitham SA, Wang K (2004) Assessment of transgenic maize events produced by particle bombardment or *Agrobacterium*-mediated transformation. *Mol Breed* 13:201–208
33. Altpeter F, Baisakh N, Beachy R, Bock R, Capell T, Christou P, Daniell H, Datta K, Dix PJ, Fauquet C, Huang N, Kohli A, Mooibroek H, Nicholson L, Nguyen TH, Nugent G, Raemakers K, Romano A, Somers DA, Stoger E, Taylor N, Visser R (2005) Particle bombardment and the genetic enhancement of crops: myths and realities. *Mol Breed* 15:305–327
34. De Neve M, De Buck S, Jacobs A, Van Montagu M, Depicker A (1997) T-DNA integration patterns in co-transformed plant cells suggest that T-DNA repeats originate from co-integration of separate T-DNAs. *Plant J* 11:15–29
35. De Buck S, Jacobs A, Van Montagu M, Depicker A (1999) The DNA sequences of T-DNA junctions suggest that complex T-DNA loci are formed by a recombination process resembling T-DNA integration. *Plant J* 20:295–304
36. Cluster PD, O'Dell M, Metzlafl M, Flavell RB (1996) Details of T-DNA structural organization from a transgenic *Petunia* population exhibiting co-suppression. *Plant Mol Biol* 32:1197–1203
37. Wolters A-MA, Trindade LM, Jacobsen E, Visser RGF (1998) Fluorescence *in situ* hybridization on extended DNA fibres as a tool to analyse complex T-DNA loci in potato. *Plant J* 13:837–847
38. Jorgensen R, Snyder C, Jones JDG (1987) T-DNA is organized predominantly in inverted repeat structures in plants transformed with *Agrobacterium tumefaciens* C58 derivatives. *Mol Gen Genet* 207:471–477
39. Krizkova L, Hroudá M (1998) Direct repeats of T-DNA integrated in tobacco chromosome: characterization of junction regions. *Plant J* 16:673–680
40. Kumar S, Fladung M (2000) Transgene repeats in aspen: molecular characterisation suggests simultaneous integration of independent T-DNAs into receptive hotspots in the host genome. *Mol Gen Genet* 264:20–28
41. Van Lijsebettens M, Inzé D, Schell J, Van Montagu M (1986) Transformed cell clones as a tool to study T-DNA integration mediated by *Agrobacterium tumefaciens*. *J Mol Biol* 188:129–145
42. Herman LMF, Van Montagu M, Depicker AG (1986) Isolation of tobacco DNA segments with plant promoter activity. *Mol Cell Biol* 6:4486–4492
43. Depicker A, Herman L, Jacobs A, Schell J, Van Montagu M (1985) Frequencies of simultaneous transformation with different T-DNAs and their relevance to the *Agrobacterium*/plant cell interaction. *Mol Gen Genet* 201:477–484
44. De Block M, Debrouwer D (1991) Two T-DNA's co-transformed into *Brassica napus* by a double *Agrobacterium tumefaciens*

- infection are mainly integrated at the same locus. *Theor Appl Genet* 82:257–263
45. Poirier Y, Ventre G, Nawrath C (2000) High-frequency linkage of co-expressing T-DNA in transgenic *Arabidopsis thaliana* transformed by vacuum-infiltration of *Agrobacterium tumefaciens*. *Theor Appl Genet* 100:487–493
 46. Radchuk VV, Van DT, Klocke E (2005) Multiple gene co-integration in *Arabidopsis thaliana* predominantly occurs in the same genetic locus after simultaneous in planta transformation with distinct *Agrobacterium tumefaciens* strains. *Plant Sci* 168:1515–1523
 47. Nam J, Matthysse AG, Gelvin SB (1997) Differences in susceptibility of *Arabidopsis* ecotypes to crown gall disease may result from a deficiency in T-DNA integration. *Plant Cell* 9:317–333
 48. Grevelding C, Fantes V, Kemper E, Schell J, Masterson R (1993) Single-copy T-DNA insertions in *Arabidopsis* are the predominant form of integration in root-derived transgenics, whereas multiple insertions are found in leaf discs. *Plant Mol Biol* 23:847–860
 49. Valvekens D, Van Montagu M, Van Lijsebettens M (1988) *Agrobacterium tumefaciens*-mediated transformation of *Arabidopsis thaliana* root explants by using kanamycin selection. *Proc Natl Acad Sci USA* 85:5536–5540
 50. Clough SJ, Bent AF (1998) Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *Plant J* 16:735–743
 51. Hobbs SLA, Warkentin TD, DeLong CMO (1993) Transgene copy number can be positively or negatively associated with transgene expression. *Plant Mol Biol* 21:17–26
 52. Shiva Prakash N, Bhojaraja R, Shivbachan SK, Hari Priya GG, Nagraj TK, Prasad V, Srikanth Babu V, Jayaprakash TL, Dasgupta S, Spencer TM, Boddupalli RS (2009) Marker-free transgenic corn plant production through co-bombardment. *Plant Cell Rep* 28:1655–1668
 53. Lowe BA, Shiva Prakash N, Way M, Mann MT, Spencer TM, Boddupalli RS (2009) Enhanced single copy integration events in corn via particle bombardment using low quantities of DNA. *Transgenic Res* 18:831–840
 54. Loc TN, Tinjuangjun P, Gatehouse AMR, Christou P, Gatehouse AJ (2002) Linear transgene constructs lacking vector backbone sequences generate transgenic rice plants which accumulate higher levels of proteins conferring insect resistance. *Mol Breed* 9:231–244
 55. Yao Q, Cong L, Chang JL, Li KX, Yang GX, He GY (2006) Low copy number gene transfer and stable expression in a commercial wheat cultivar via particle bombardment. *J Exp Bot* 57:3737–3746
 56. Fu X, Duc LT, Fontana S, Bong BB, Tinjuangjun P, Sudhakar D, Twyman RM, Christou P, Kohli A (2000) Linear transgene constructs lacking vector backbone sequences generate low-copy-number transgenic plants with simple integration patterns. *Transgenic Res* 9:11–19
 57. Odell JT, Nagy F, Chua N-H (1985) Identification of DNA sequences required for activity of the cauliflower mosaic virus 35S promoter. *Nature* 313:810–812
 58. Fang R-X, Nagy F, Sivasubramaniam S, Chua N-H (1989) Multiple *cis* regulatory elements for maximal expression of the cauliflower mosaic virus 35S promoter in transgenic plants. *Plant Cell* 1:141–150
 59. Jorgensen RA, Cluster PD, English J, Que Q, Napoli CA (1996) Chalcone synthase cosuppression phenotypes in petunia flowers: comparison of sense vs. antisense constructs and single-copy vs. complex T-DNA sequences. *Plant Mol Biol* 31:957–973
 60. Elmayan T, Vaucheret H (1996) Expression of single copies of a strongly expressed 35S transgene can be silenced post-transcriptionally. *Plant J* 9:787–797
 61. Que Q, Wang H-Y, English JJ, Jorgensen RA (1997) The frequency and degree of cosuppression by sense chalcone synthase transgenes are dependent on transgene promoter strength and are reduced by premature nonsense codons in the transgene coding sequence. *Plant Cell* 9:1357–1368
 62. Lechtenberg B, Schubert D, Forsbach A, Gils M, Schmidt R (2003) Neither inverted repeat T-DNA configurations nor arrangements of tandemly repeated transgenes are sufficient to trigger transgene silencing. *Plant J* 34:507–517
 63. Marjanac G, Karimi M, Naudts M, Beeckman T, Depicker A, De Buck S (2009) Gene silencing induced by hairpin or inverted repeated sense transgenes varies among promoters and cell types. *New Phytol* 184:851–864
 64. Chuang C-F, Meyerowitz EM (2000) Specific and heritable genetic interference by double-stranded RNA in *Arabidopsis thaliana*. *Proc Natl Acad Sci USA* 97:4985–4990
 65. Vaucheret H, Elmayan T, Thierry D, van der Geest A, Hall T, Conner AJ, Mlynarova L, Nap J-P (1998) Flank matrix attachment regions (MARs) from chicken, bean, yeast or tobacco do not prevent homology-dependent *trans*-silencing in transgenic tobacco plants. *Mol Gen Genet* 259:388–392
 66. De Bolle MFC, Butaye KMJ, Coucke WJW, Goderis IJWM, Wouters PFJ, van Boxel N, Broekaert WF, Cammue BPA (2003) Analysis of the influence of promoter elements and a matrix attachment region on the inter-individual variation of transgene expression in populations of *Arabidopsis thaliana*. *Plant Sci* 165:169–179
 67. Mathé C, Peresetsky A, Déhais P, Van Montagu M, Rouzé P (1999) Classification of *Arabidopsis thaliana* gene sequences: clustering of coding sequences into two groups according to codon usage improves gene prediction. *J Mol Biol* 285:1977–1991
 68. Rouwendal GJA, Mendes O, Wolbert EJH, de Boer AD (1997) Enhanced expression in tobacco of the gene encoding green fluorescent protein by modification of its codon usage. *Plant Mol Biol* 33:989–999
 69. Schubert D, Lechtenberg B, Forsbach A, Gils M, Bahadur S, Schmidt R (2004) Silencing in *Arabidopsis* T-DNA transformants: the predominant role of a gene-specific RNA sensing mechanism versus position effects. *Plant Cell* 16:2561–2572

70. De Jaeger G, Scheffer S, Jacobs A, Zambre M, Zobell O, Goossens A, Depicker A, Angenon G (2002) Boosting heterologous protein production in transgenic dicotyledonous seeds using *Phaseolus vulgaris* regulatory sequences. *Nat Biotechnol* 20:1265–1268
71. Mascarenhas D, Mettler IJ, Pierce DA, Lowe HW (1990) Intron-mediated enhancement of heterologous gene expression in maize. *Plant Mol Biol* 15:913–920
72. Bourdon V, Harvey A, Lonsdale DM (2001) Introns and their positions affect the translational activity of mRNA in plant cells. *EMBO Rep* 2:394–398
73. Rose AB, Elfersi T, Parra G, Korf I (2008) Promoter-proximal introns in *Arabidopsis thaliana* are enriched in dispersed signals that elevate gene expression. *Plant Cell* 20:543–551
74. Barlett JG, Snape JW, Harwood WA (2009) Intron-mediated enhancement as a method for increasing transgene expression levels in barley. *Plant Biotechnol J* 7:856–866
75. Rose AB (2004) The effect of intron location on intron-mediated enhancement of gene expression in *Arabidopsis*. *Plant J* 40:744–751
76. Tanaka A, Mita S, Ohta S, Kyojuka J, Shimamoto K, Nakamura K (1990) Enhancement of foreign gene expression by a dicot intron in rice but not in tobacco is correlated with an increased level of mRNA and an efficient splicing of the intron. *Nucleic Acids Res* 18:6767–6770
77. Allen GC, Spiker S, Thompson WF (2000) Use of matrix attachment regions (MARs) to minimize transgene silencing. *Plant Mol Biol* 43:361–376
78. De Bolle MFC, Butaye KMJ, Goderis IJWM, Wouters PFJ, Jacobs A, Delauré SL, Depicker A, Cammue BPA (2007) The influence of matrix attachment regions on transgene expression in *Arabidopsis thaliana* wild type and gene silencing mutants. *Plant Mol Biol* 63:533–543
79. Butaye KMJ, Cammue BPA, Delauré SL, De Bolle MFC (2005) Approaches to minimize variation of transgene expression in plants. *Mol Breed* 16:79–91
80. Petersen K, Leah R, Knudsen S, Cameron-Mills V (2002) Matrix attachment regions (MARs) enhance transformation frequencies and reduce variance of transgene expression of barley. *Plant Mol Biol* 49:45–58
81. Ülker B, Allen GC, Thompson WF, Spiker S, Weissinger AK (1999) A tobacco matrix attachment region reduces the loss of transgene expression in the progeny of transgenic tobacco plants. *Plant J* 18:253–263
82. Vain P, Worland B, Kohli A, Snape JW, Christou P, Allen GC, Thompson WF (1999) Matrix attachment regions increase transgene expression levels and stability in transgenic rice plants and their progeny. *Plant J* 18:233–242
83. Mlynárová L, Jansen RC, Conner AJ, Stiekema WJ, Nap J-P (1995) The MAR-mediated reduction in position effect can be uncoupled from copy number-dependent expression in transgenic plants. *Plant Cell* 7:599–609
84. Sels J, Delauré SL, Aerts AM, Proost P, Cammue BPA, De Bolle MFC (2007) Use of a PTGS-MAR expression system for efficient in planta production of bioactive *Arabidopsis thaliana* plant defensins. *Transgenic Res* 16:531–538
85. Iglesias VA, Moscone EA, Papp I, Neuhauser F, Michalowski S, Phelan T, Spiker S, Matzke M, Matzke AJM (1997) Molecular and cytogenetic analyses of stably and unstably expressed transgene loci in tobacco. *Plant Cell* 9:1251–1264
86. Day CD, Lee E, Kobayashi J, Holappa LD, Albert H, Ow DW (2000) Transgene integration into the same chromosome location can produce alleles that express at a predictable level, or alleles that are differentially silenced. *Genes Dev* 14:2869–2880
87. De Wilde C, Podevin N, Windels P, Depicker A (2001) Silencing of antibody genes in plants with single-copy transgene inserts as a result of gene dosage effects. *Mol Genet Genomics* 265:647–653
88. De Paepe A, De Buck S, Hoorelbeke K, Nolf J, Peck I, Depicker A (2009) High frequency of single-copy T-DNA transformants by floral dip in *CRE*-expressing *Arabidopsis* plants. *Plant J* 59:517–527
89. Meyer P, Heidmann I (1994) Epigenetic variants of a transgenic petunia line show hypermethylation in transgene DNA: an indication for specific recognition of foreign DNA in transgenic plants. *Mol Gen Genet* 243:390–399
90. Elomaa P, Helariutta Y, Griesbach RJ, Kotilainen M, Seppänen P, Teeri TH (1995) Transgene inactivation in *Petunia hybrida* is influenced by the properties of the foreign gene. *Mol Gen Genet* 248:649–656
91. Peach C, Velten J (1991) Transgene expression variability (position effect) of CAT and GUS reporter genes driven by linked divergent T-DNA promoters. *Plant Mol Biol* 17:49–60
92. Meyer P, Heidmann I, Niedenhof I (1993) Differences in DNA-methylation are associated with a paramutation phenomenon in transgenic petunia. *Plant J* 4:89–100
93. Holmes-Davis R, Comai L (1998) Nuclear matrix attachment regions and plant gene expression. *Trends Plant Sci* 3:91–97
94. Meza TJ, Stangeland B, Mercy IS, Skårn M, Nymoen DA, Berg A, Butenko MA, Häkelien A-M, Haslekås C, Meza-Zepeda LA, Aalen RB (2002) Analyses of single-copy *Arabidopsis* T-DNA-transformed lines show that the presence of vector backbone sequences, short inverted repeats and DNA methylation is not sufficient or necessary for the induction of transgene silencing. *Nucleic Acids Res* 30:4556–4566
95. Kunz C, Narangajavana J, Jakowitsch J, Park Y-D, Delon TR, Kovarik A, Koukaková B, van der Winden J, Moscone E, Aufsatz W, Mette MF, Matzke M, Matzke AJM (2003) Studies on the effects of a flanking repetitive sequence on the expression of single-copy transgenes in *Nicotiana glauca* and in *N. sylvestris*-*N. tomentosiformis* hybrids. *Plant Mol Biol* 52:203–215

96. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, Brandon RC, Rogers Y-HC, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C, Baxter EG, Helt G, Nelson CR, Gabor Miklos GL, Abril JF, Agbayani A, An H-J, Andrews-Pfannkoch C, Baldwin D, Ballew RM, Basu A, Baxendale J, Bayraktaroglu L, Beasley EM, Beeson KY, Benos PV, Berman BP, Bhandari D, Bolshakov S, Borkova D, Botchan MR, Bouck J, Brokstein P, Brottier P, Burtis KC, Busam DA, Butler H, Cadieu E, Center A, Chandra I, Cherry JM, Cawley S, Dahlke C, Davenport LB, Davies P, de Pablos B, Delcher A, Deng Z, Deslattes May A, Dew I, Dietz SM, Dodson K, Doup LE, Downes M, Dugan-Rocha S, Dunkov BC, Dunn P, Durbin KJ, Evangelista CC, Ferraz C, Ferriera S, Fleischmann W, Fosler C, Gabrielian AE, Garg NS, Gelbart WM, Glasser K, Glodek A, Gong F, Gorrell JH, Gu Z, Guan P, Harris M, Harris NL, Harvey D, Heiman TJ, Hernandez JR, Houck J, Hostin D, Houston KA, Howland TJ, Wei M-H, Ibegwam X, Jalali M, Kalush F, Karpen GH, Ke Z, Kennison JA, Ketchum KA, Kimmel BE, Kodira CD, Kraft C, Kravitz S, Kulp D, Lai Z, Lasko P, Lei Y, Levitsky AA, Li J, Li Z, Liang Y, Lin X, Liu X, Mattei B, McIntosh TC, McLeod MP, McPherson D, Merkulov G, Milshina NV, Mobarry C, Morris J, Moshrefi A, Mount SM, Moy M, Murphy B, Murphy L, Muzny DM, Nelson DL, Nelson DR, Nelson KA, Nixon K, Nusskern DR, Pacleb JM, Palazzolo M, Pittman GS, Pan S, Pollard J, Puri V, Reese MG, Reinert K, Remington K, Saunders RDC, Scheeler F, Shen H, Shue BC, Sidén-Kiamos I, Simpson M, Skupski MP, Smith T, Spier E, Spradling AC, Stapleton M, Strong R, Sun E, Svirskas R, Tector C, Turner R, Venter E, Wang AH, Wang X, Wang Z-Y, Wassarman DA, Weinstock GM, Weissenbach J, Williams SM, Woodage T, Worley KC, Wu D, Yang S, Yao QA, Ye J, Yeh R-F, Zaveri JS, Zhan M, Zhang G, Zhao Q, Zheng L, Zheng XH, Zhong FN, Zhong W, Zhou X, Zhu S, Zhu X, Smith HO, Gibbs RA, Myers EW, Rubin GM, Venter JC (2000) The genome sequence of *Drosophila melanogaster*. *Science* 287:2185–2195
97. Puchta H (2002) Gene replacement by homologous recombination in plants. *Plant Mol Biol* 48:173–182
98. Hanin M, Paszkowski J (2003) Plant genome modification by homologous recombination. *Curr Opin Plant Biol* 6:157–162
99. Hanin M, Volrath S, Bogucki A, Briker M, Ward E, Paszkowski J (2001) Gene targeting in *Arabidopsis*. *Plant J* 28:671–677
100. Shaked H, Melamed-Bessudo C, Levy AA (2005) High-frequency gene targeting in *Arabidopsis* plants expressing the yeast *RAD54* gene. *Proc Natl Acad Sci USA* 102:12265–12269
101. Terada R, Johzuka-Hisatomi Y, Saitoh M, Asao H, Iida S (2007) Gene targeting by homologous recombination as a biotechnological tool for rice functional genomics. *Plant Physiol* 144:846–856
102. Salomon S, Puchta H (1998) Capture of genomic and T-DNA sequences during double-strand break repair in somatic plant cells. *EMBO J* 17:6086–6095
103. Tzfira T, Frankman LR, Vaidya M, Citovsky V (2003) Site-specific integration of *Agrobacterium tumefaciens* T-DNA via double-stranded intermediates. *Plant Physiol* 133:1011–1023
104. Lloyd A, Plaisier CL, Carroll D, Drews GN (2005) Targeted mutagenesis using zinc-finger nucleases in *Arabidopsis*. *Proc Natl Acad Sci USA* 102:2232–2237
105. Wright DA, Townsend JA, Winfrey RJ Jr, Irwin PA, Rajagopal J, Lonosky PM, Hall BD, Jondle MD, Voytas DF (2005) High-frequency homologous recombination in plants mediated by zinc-finger nucleases. *Plant J* 44:693–705
106. Cai CQ, Doyon Y, Ainley WM, Miller JC, DeKelver RC, Moehle EA, Rock JM, Lee Y-L, Garrison R, Schulenberg L, Blue R, Worden A, Baker L, Faraji F, Zhang L, Holmes MC, Rebar EJ, Collingwood TN, Rubin-Wilson B, Gregory PD, Urnov FD, Petolino JF (2009) Targeted transgene integration in plant cells using designed zinc finger nucleases. *Plant Mol Biol* 69:699–709
107. Carroll D, Morton JJ, Beumer KJ, Segal DJ (2006) Design, construction and *in vitro* testing of zinc finger nucleases. *Nat Protoc* 1:1329–1341
108. De Buck S, Van Montagu M, Depicker A (2001) Transgene silencing of invertedly repeated transgenes is released upon deletion of one of the transgenes involved. *Plant Mol Biol* 46:433–445
109. Hobbs SLA, Kpodar P, DeLong CMO (1990) The effect of T-DNA copy number, position and methylation on reporter gene expression in tobacco transformants. *Plant Mol Biol* 15:851–864
110. Assaad FF, Tucker KL, Signer ER (1993) Epigenetic repeat-induced gene silencing (RIGS) in *Arabidopsis*. *Plant Mol Biol* 22:1067–1085
111. Sijen T, Wellink J, Hiriart J-B, van Kammen A (1996) RNA-mediated virus resistance: role of repeated transgenes and delineation of targeted regions. *Plant Cell* 8:2277–2294
112. de Carvalho F, Gheysen G, Kushnir S, Van Montagu M, Inzé D, Castresana C (1992) Suppression of β -1, 3-glucanase transgene expression in homozygous plants. *EMBO J* 11:2595–2602
113. De Neve M, De Buck S, De Wilde C, Van Houdt H, Strobbe I, Jacobs A, Van Montagu M, Depicker A (1999) Gene silencing results in instability of antibody production in transgenic plants. *Mol Gen Genet* 260:582–592
114. Ingelbrecht I, Van Houdt H, Van Montagu M, Depicker A (1994) Posttranscriptional silencing of reporter transgenes in tobacco correlates with DNA methylation. *Proc Natl Acad Sci USA* 91:10502–10506
115. Stam M, de Bruin R, Kenter S, van der Hoorn RAL, van Blokland R, Mol JNM, Kooter JM (1997) Post-transcriptional silencing of chalcone synthase in *Petunia* by inverted transgene repeats. *Plant J* 12:63–82
116. Jones JDG, Gilbert DE, Grady KL, Jorgensen RA (1987) T-DNA structure and gene expression in petunia plants transformed by *Agrobacterium tumefaciens* C58 derivatives. *Mol Gen Genet* 207:478–485

117. Van Houdt H, Kovařík A, Van Montagu M, Depicker A (2000) Cross-talk between posttranscriptionally silenced neomycin phosphotransferase II transgenes. *FEBS Lett* 467:41–46
118. De Buck S, Depicker A (2001) Disruption of their palindromic arrangement leads to selective loss of DNA methylation in inversely repeated *gus* transgenes in *Arabidopsis*. *Mol Genet Genomics* 265:1060–1068
119. Vaucheret H (1993) Identification of a general silencer for 19S and 35S promoters in a transgenic tobacco plant: 90 bp of homology in the promoter sequence are sufficient for trans-inactivation. *CR Acad Sci Paris, Life Sci* 316:1471–1483
120. Hutvágner G, Mlynárová L, Nap J-P (2000) Detailed characterization of the posttranscriptional gene-silencing-related small RNA in a GUS gene-silenced tobacco. *RNA* 6:1445–1454
121. Matzke AJM, Neuhuber F, Park Y-D, Ambros PF, Matzke MA (1994) Homology-dependent gene silencing in transgenic plants: epistatic silencing loci contain multiple copies of methylated transgenes. *Mol Gen Genet* 244:219–229
122. Hagan ND, Spencer D, Moore AE, Higgins TJV (2003) Changes in methylation during progressive transcriptional silencing in transgenic subterranean clover. *Plant Biotechnol J* 1:479–490
123. Meza TJ, Enerly E, Børud B, Larsen F, Mandal A, Aalen RB, Jakobsen KS (2002) A human CpG island randomly inserted into a plant genome is protected from methylation. *Transgenic Res* 11:133–142
124. Fischer U, Kuhlmann M, Pecinka A, Schmidt R, Mette MF (2008) Local DNA features affect RNA-directed transcriptional gene silencing and DNA methylation. *Plant J* 53:1–10
125. Brunner AM, Li J, DiFazio SP, Shevchenko O, Montgomery BE, Mohamed R, Wei H, Ma C, Elias AA, VanWormer K, Strauss SH (2007) Genetic containment of forest plantations. *Tree Genet Genomes* 3:75–100
126. De Wilde C, Van Houdt H, De Buck S, Angenon G, De Jaeger G, Depicker A (2000) Plants as bioreactors for protein production: avoiding the problem of transgene silencing. *Plant Mol Biol* 43:347–359
127. Hamilton AJ, Brown S, Yuanhai H, Ishizuka M, Lowe A, Alpuche Solis A-G, Grierson D (1998) A transgene with repeated DNA causes high frequency, post-transcriptional suppression of ACC-oxidase gene expression in tomato. *Plant J* 15:737–746
128. De Buck S, Depicker A (2004) Gene expression and level of expression. In: Christou P, Klee H (eds) *Handbook of plant biotechnology*, vol 1. Wiley, Chichester, pp 331–345
129. De Buck S, Peck I, De Wilde C, Marjanac G, Nolf J, De Paepe A, Depicker A (2007) Generation of single-copy T-DNA transformants in *Arabidopsis* by the CRE/*loxP* recombination-mediated resolution system. *Plant Physiol* 145:1171–1182
130. Marjanac G, De Paepe A, Peck I, Jacobs A, De Buck S, Depicker A (2008) Evaluation of CRE-mediated excision approaches in *Arabidopsis thaliana*. *Transgenic Res* 17:239–250
131. Coppoolse ER, de Vroomen MJ, Roelofs D, Smit J, van Gennip F, Hersmus BJM, Nijkamp HJJ, van Haaren MJJ (2003) Cre recombinase expression can result in phenotypic aberrations in plants. *Plant Mol Biol* 51:263–279
132. Vergunst AC, Schrammeijer B, den Dulk-Ras A, de Vlaam CMT, Regensburg-Tuink TJG, Hooykaas PJJ (2000) VirB/D4-dependent protein translocation from *Agrobacterium* into plant cells. *Science* 290:979–982
133. Verweire D, Verleyen K, De Buck S, Claeys M, Angenon G (2007) Marker-free transgenic plants through genetically programmed auto-excision. *Plant Physiol* 145:1220–1231
134. Srivastava V, Anderson OD, Ow DW (1999) Single-copy transgenic wheat generated through the resolution of complex integration patterns. *Proc Natl Acad Sci USA* 96:11117–11121
135. Srivastava V, Ow DW (2001) Single-copy primary transformants of maize obtained through the co-introduction of a recombinase-expressing construct. *Plant Mol Biol* 46:561–566
136. Kilby NJ, Davies GJ, Snaith MR, Murray JAH (1995) FLP recombinase in transgenic plants: constitutive activity in stably transformed tobacco and generation of marked cell clones in *Arabidopsis*. *Plant J* 8:637–652
137. Kumar S, Thompson WF (2009) Simultaneous excision of two transgene flanking sequences and resolution of complex integration loci. *Plant Mol Biol* 69:23–32
138. Vergunst AC, Jansen LET, Hooykaas PJJ (1998) Site-specific integration of *Agrobacterium* T-DNA in *Arabidopsis thaliana* mediated by Cre recombinase. *Nucleic Acids Res* 26: 2729–2734
139. Srivastava V, Ow DW (2001) Biolistic mediated site-specific integration in rice. *Mol Breed* 8:345–349
140. Srivastava V, Ow DW (2004) Marker-free site-specific gene integration in plants. *Trends Biotechnol* 22:627–629
141. Chawla R, Ariza-Nieto M, Wilson AJ, Moore SK, Srivastava V (2006) Transgene expression produced by biolistic-mediated, site-specific gene integration is consistently inherited by the subsequent generations. *Plant Biotechnol J* 4:209–218
142. Nanto K, Sato K, Katayama Y, Ebinuma H (2009) Expression of a transgene exchanged by the recombinase-mediated cassette exchange (RMCE) method in plants. *Plant Cell Rep* 28:777–785
143. Li Z, Xing A, Moon BP, McCardell RP, Mills K, Falco SC (2009) Site-specific integration of transgenes in soybean via recombinase-mediated DNA cassette exchange. *Plant Physiol* 151:1087–1095
144. Hansen G, Chilton M-D (1996) "Agrolistic" transformation of plant cells: integration of T-strands generated *in planta*. *Proc Natl Acad Sci USA* 93:14978–14983
145. Hansen G, Shillito RD, Chilton M-D (1997) T-strand integration in maize protoplasts after codelivery of a T-DNA substrate and virulence genes. *Proc Natl Acad Sci USA* 94:11726–11730

146. De Block M, Debrouwer D, Moens T (1997) The development of a nuclear male sterility system in wheat. Expression of the barnase gene under the control of tapetum specific promoters. *Theor Appl Genet* 95:125–131
147. Dalmay T, Hamilton A, Rudd S, Angell S, Baulcombe DC (2000) An RNA-dependent RNA polymerase gene in *Arabidopsis* is required for posttranscriptional gene silencing mediated by a transgene but not by a virus. *Cell* 101:543–553
148. Mourrain P, Béclin C, Elmayan T, Feuerbach F, Godon C, Morel J-B, Jouette D, Lacombe A-M, Nikic S, Picault N, Rémoûé K, Sanial M, Vo T-A, Vaucheret H (2000) *Arabidopsis* SGS2 and SGS3 genes are required for posttranscriptional gene silencing and natural virus resistance. *Cell* 101:533–542
149. Verdaguer B, de Kochko A, Beachy RN, Fauquet C (1996) Isolation and expression in transgenic tobacco and rice plants of the cassava vein mosaic virus (CVMV) promoter. *Plant Mol Biol* 31:1129–1139
150. Goossens A, Dillen W, De Clercq J, Van Montagu M, Angenon G (1999) The *arcelin-5* gene of *Phaseolus vulgaris* directs high seed-specific expression in transgenic *Phaseolus acutifolius* and *Arabidopsis* plants. *Plant Physiol* 120:1095–1104

Transgenic Crops Resistant to Fungal, Bacterial and Viral Pathogens

TENNANT PAULA¹, FUCHS MARC²

¹Department of Life Sciences, University of the West Indies, Mona, Kingston, WI, Jamaica

²Department of Plant Pathology and Plant-Microbe Biology, Cornell University, New York State Agricultural Experiment Station, Geneva, NY, USA

Article Outline

Glossary

Definition of the Subject

Introduction

Transgenic Resistance to Fungal Pathogens

Transgenic Resistance to Bacterial Pathogens

Transgenic Resistance to Viral Pathogens

Status of Commercialized Disease Resistant Transgenic Crops

Conclusions

Future Directions

Bibliography

Glossary

Bacterium A unicellular, prokaryotic microorganism with a wide range of shapes that is surrounded by a lipid membrane and a cell wall made of peptidoglycan, and has few intracellular structures, no nucleus but a single circular chromosome located in the cytoplasm and reproduces by binary fission.

Fungus A eukaryotic organism such as a yeast or a mold, with cell walls that contain glucans and chitin, membrane-bound nuclei with chromosomes and cytoplasmic organelles, soluble carbohydrates, and storage compounds, and exhibits enormous diversity in life cycle strategies, morphologies, and ecologies.

Pathogen An infectious agent that causes disease to its host.

Pathogenicity The capacity of a pathogen to cause disease.

Pathogenicity factors Products encoded by a pathogen that are crucial for the establishment of disease, e.g., proteins involved in the attachment of the pathogen to the plant surface or penetration and colonization of the host.

Resistance The reaction of a host to a pathogen that derives from preformed defenses and/or defense responses induced following infection.

Systemic acquired resistance (SAR) A mechanism of induced defense that acts nonspecifically throughout the plant, is associated with the accumulation of pathogenesis-related proteins and requires the signal molecule, salicylic acid.

Transgenic plant A plant that results from the application of recombinant DNA and plant tissue culture technologies.

Virus A small obligate parasite that utilizes the cellular metabolic pathways of its host organism to replicate its genes which are made up of deoxyribonucleic or ribonucleic acid and are encapsidated in a protein coat.

Definition of the Subject

Loss of crops due to fungal, bacterial, and viral diseases can have a large impact on human food supplies and local economies, as well as on the social stability of rural communities. It is conservatively estimated that diseases, insects, and weeds together cause 30–40% loss of all crops worldwide [1]. Annual losses worldwide

due to plant diseases are estimated to ~14% of total losses and about \$220 billion. In addition, the need for measures to control diseases limits the acreage of land available for cultivation, restricts the crops that can be grown in fields already contaminated with certain pathogens and necessitates the use of agrochemicals for treating seeds, fumigating soils, spraying plants, and applying fruit postharvest treatments. Such control measures add to the cost of food production and toxic chemicals can be harmful to human health and the environment [1].

Since the beginning of agriculture, crop plants have been domesticated and improved for various traits such as increased yield, improved nutritious composition, enhanced disease resistance, and tolerance to abiotic factors, among others. The selection and use of resistant crop cultivars is the most efficient and most environmentally friendly strategy to mitigate the impact of fungal, bacterial, and viral diseases. Resistance to pathogens can be achieved by application of disease-suppressing cultural practices, use of plant defense-promoting substances, deployment of biological agents antagonistic to the pathogens that cause disease, agrochemicals, conventional breeding strategies, and genetic engineering [1]. The need for controlling plant diseases effectively is not only a major challenge, but also a necessity to reduce food losses while improving food quality and safeguarding the environment.

The advent of molecular genetics and plant transformation, in combination with an increased understanding of pathogen–host interactions, has opened new avenues for the development of disease-resistant crops through genetic engineering. Research groups from academia and industry created the first transgenic plants in the early 1980s [2–5]. These early transgenic plants were laboratory specimens (tobacco, petunia, and sunflower). Subsequent research developed transgenic plants, including horticultural crops, with useful traits such as disease resistance. Transgenic crops resistant to fungal, bacterial, and viral diseases were created following extensive research. The first disease-resistant transgenic crop, i.e., a virus-resistant transgenic summer squash, was deregulated in the USA in 1994. Several cultivars derived from the deregulated transgenic summer squash lines were commercialized in the USA in 1996 [6]. The adoption rate of virus-resistant transgenic summer squash has been increasing steadily since

the initial releases. Other commercially available virus-resistant transgenic crops include papaya in the USA, and sweet pepper and tomato in the People's Republic of China. Transgenic crops resistant to fungal and bacterial diseases have been developed and evaluated under field conditions. However, in spite of remarkable progress, their development and release lags behind virus-resistant transgenic crops.

Introduction

Plants provide microorganisms, including viruses, with a diversity of habitats [7]. Ensuing interactions are either detrimental or beneficial and are classified as neutralism, commensalism, synergism, mutualism, amensalism, or parasitism. It is the latter association that has been a major challenge to crop production for centuries.

In the parasitic interaction, an organism lives within the plant host or on its outer surfaces and obtains its nutrients from the latter [1]. The removal of nutrients often affects normal growth and development of the plant and may be associated with pathogenicity. Virus infections commonly result in the inhibition of plant growth as well as the development of undesirable morphological changes and decreased yield [8]. Unlike fungal and bacterial pathogens, viruses induce disease in plant hosts primarily by utilizing cellular components and disrupting cellular processes. Fungi are responsible for diseases referred to as blight, gray mold, bunts, powdery, and downy mildews. These microbes have evolved at least three strategies that lead to less-than-healthy plants; enzymes that breakdown the plant cell wall and cuticle, toxins that either reduce the activity of the host cells or inhibit it completely, and plant-specific hormones that interfere with the hormonal equilibrium of the plant cell [9]. Bacterial pathogens cause pre- and post-harvest diseases that are characterized by cankers, galls, wilts, spotting of leaves, and stem or fruit rots of various vegetable and fruit crops. The invasion of healthy plant tissue through a wound or an area of dead or dying tissue is generally followed by the initiation of infection, which is similar to that of fungi, and is associated with the breakdown of cell walls, the production of toxins, exo-polysaccharides or proteins that mimic plant hormones.

Plant diseases, whether initiated by viruses, bacteria, or fungi, are of paramount importance because of the destruction they cause to plants and plant products. The effects not only threaten crop productivity and reduce farmers' net income, but also the food supply, and by extension, the economies of rural areas and even countries [10]. Zadoks and Schein [11] differentiate between primary and secondary losses and direct and indirect losses. Primary losses refer to those sustained because of plant disease (before or after harvest) in yield or wages, while secondary losses are losses of future production capacity. On the other hand, losses in quality, quantity, and production capacity, and socioeconomic losses are classified as direct and indirect losses, respectively.

Disease control efforts to date have focused on various crop husbandry techniques including crop rotation, and the avoidance of spread of infested soil and pathogen-carrying plant materials. The most widely used management strategy is the application of agrochemicals, particularly against fungal pathogens. Chemical protection against bacterial pathogens is not as evolved as that against fungi, and is not applicable to viruses. The most widely used methods against bacterial diseases primarily center on phytosanitary practices, harvesting techniques, and storage conditions, and eventually on the use of antibiotics. However, the intensive use of agrochemicals has led to the development of resistance in various microbial populations and in some cases the chemicals are no longer used because of their high toxicity or problems with persistence in the soil. Nowadays, the focus is on strategies that allow for crop production with minimal use of agrochemicals. The most effective strategy to mitigate the impact of diseases on agriculture involves breeding-resistant crop cultivars. The majority of the cultivars available today have been generated through the classical method of hybridization [12] and the introgression of genes responsible for resistance. Resistant plants ward off pathogen attack by an innate immunity of each cell, systemic signals emanating from infection sites and a suite of cellular responses that follow the activation of the gene for resistance. However, there are many pathogens for which no effective sources of disease resistance have been identified. Technologies such as genetic engineering and gene transfer methods offer an alternative means for the development of disease

resistant cultivars. In contrast to conventional breeding which involves the random mixing of the tens of thousands of genes present in both the resistant and susceptible plant, the latter technology allows the transfer of only the resistance gene to the susceptible plant and the preservation of its valuable economic traits. Moreover, the genetic sources for disease resistance are not limited to closely related plant species. During the last decade, considerable progress has been made in the development of virus resistant transgenic plants, while the development of fungal- and bacterial-resistant transgenic plants has lagged behind.

This chapter examines the different strategies devised for the development of resistance to fungal, bacterial, and viral diseases through genetic engineering. A description of the genes employed to confer resistance against given pathogens and the mechanisms underlying engineered resistance are provided. Emphasis will be on horticultural crops and the recent applications of transgenic resistance to agriculture. Factors affecting the development, release, and adoption of transgenic crops with resistance to fungal, bacterial, and viral diseases will be identified and discussed. The final sections offer overall conclusions, cover the impending direction of the development of transgenic disease resistance and some reflection on the impact of transgenic crops resistant to fungi, bacteria, and viruses in terms of effective disease management, mitigation of disease impact on agriculture, improved production of quality food, and better preservation of the environment.

Transgenic Resistance to Fungal Pathogens

Plants have evolved a complex pathogen defense system that is driven by tight balances between the induction of specific defense pathways and cell death. Much progress has recently been made in identifying and understanding a number of the genes and gene products involved in these defenses as well as the signaling pathways. The findings have prompted the development of a variety of strategies for producing fungus-resistant transgenic plants that are based on the manipulation of the plant's innate defense system, the production of antifungal proteins or the neutralization of fungal pathogenicity factors.

The earliest strategy employed in the development of resistance against fungal pathogens involved the

enhancement of the plant's natural defenses through the overexpression of pathogenesis-related proteins (PR- proteins). PR- proteins form part of the general resistance mechanism in plants, that is, systemic acquired resistance (SAR), which is induced by necrotizing pathogens or by treatment with chemicals such as salicylic acid. Initial studies on the development of fungus-resistant transgenic plants examined the effects of expressing a single PR-protein gene in transgenic plants, while later studies looked at the co-expression of combinations thereof, in an attempt to improve on the levels of resistance, to target a wider spectrum of fungal pathogens and to provide longer-term protection.

In 1991, Broglie and coworkers [13] reported on the first fungus-resistant transgenic plants. Transgenic tobacco and rapeseed plants, transformed with a chitinase gene isolated from bean, exhibited enhanced resistance against the soil-borne fungus, *Rhizoctonia solani*. Presumably, expression of the chitinase transgene played a dual role in protecting against infection; that is, the enzyme inhibited growth of the pathogen by cell wall digestion and the released pathogen-borne elicitors induced additional defense reactions in the plants. Following the first pioneering work of Broglie et al. [13], Lorito et al. [14] showed that an endochitinase gene (*ThEn-42*) from the mycoparasitic fungus, *Trichoderma harzianum*, conferred similar resistance to the fungal pathogens, *Alternaria alternata*, *A. solani*, *Botrytis cinerea*, and *Rhizoctonia solani* in transgenic tobacco and potato (cv. 'Desiree'). The effectiveness of glucanase, another PR-protein, in transgenic plants was demonstrated in banana transformed with β -1, 3 glucanase from soybean. Increased protection was observed when transgenic banana was infected with *Fusarium oxysporum* f. sp. *cubense* [15]. Further, the combinatorial expression of two chitinase genes [16] or chitinase and glucanase genes was found to give increased levels of protection than the expression of the PR-proteins in single transformants [17]. Subsequent strategies employed the co-expression of transgenes encoding differently acting antifungal proteins with similar results. For example, ÓBrian et al. [18] confirmed protection against *Rhizoctonia solani* in transgenic tobacco carrying three barley proteins; chitinase, β -1, 3 glucanase, and plant ribosome-inactivating protein (RIP). While

the first two enzymes targeted the cell wall of the pathogen, RIP targeted and inactivated fungal ribosomes. RIPs possess N-glycosidase activity and catalyze the removal of an adenine residue from the 28S rRNA, thus inhibiting protein elongation.

Another emerging strategy against fungal pathogens involves the use of genes encoding small cysteine-rich antifungal proteins. These proteins play a role in the plant's innate defense mechanism and include thionins, defensins, and lipid transfer proteins.

The antimicrobial activities of thionins are well known, particularly those found in barley. Antimicrobial action, most likely based on the induction of membrane permeabilization resulting in cell disruption and death, is not limited to fungi and these proteins are effective against bacteria and oomycetes. Given this growth inhibition activity, transgenic plants overexpressing thionins have been generated to provide fungal resistance. Epplé et al. [19] overexpressed an endogenous endothionin *Thi2.1* gene in *Arabidopsis thaliana* and observed enhanced resistance against *Fusarium oxysporum* f. sp. *matthiolae*. Similar levels of resistance were obtained with *Thi2.1* transgenic tomato against bacterial and *Fusarium* wilt [20].

Other small cysteine-rich proteins referred to as defensins are increasingly becoming recognized as candidates of defense transgenes. Defensins were originally called γ -thionins due to similarities to *a*- and *b*-thionins, but it was later shown that the former proteins shared appreciable structural and functional commonalities to insect and mammalian defensins, hence the name change. As with thionins, the mechanism of action has yet to be confirmed. However there is strong evidence of the induction of ion fluxes across the plasma membranes of living fungal hyphae followed by interactions with glycosylceramides at the fungal cell surface. Expression of single-protein defensin gene constructs has been deployed in transgenic plants as well as cleavable, chimeric polyprotein constructs in an attempt to increase expression levels in planta and to introduce broad-spectrum resistance. The constitutive expression of a mustard defensin conferred resistance to *Fusarium moniliforme* and *Phytophthora parasitica* pv. *nicotianae* in tobacco and *Pheoisariopsis personata* and *Cercospora arachidicola*, which jointly cause late leaf spot disease in peanuts [21]. François et al. [22] engineered a polyprotein construct using defensin genes *DmAMP1*

from *Dahlia merckii* seeds and *Rs-AFP2* from *Raphanus sativus* seeds. The genes were joined by a linker peptide obtained from *Impatiens balsamina* seeds [22] and used for the transformation of *Arabidopsis*. Antifungal activity was obtained against *Fusarium culmorum* in in vitro assays using purified fractions from extracellular fluids of the transgenic plants.

Lipid transfer proteins (LTPs) are the third group of small cysteine-rich antifungal proteins (ca. 10 KDa) found in higher plants. As the name implies, LTPs are involved in shuttling phospholipids and other fatty acid groups between cell membranes and bind acyl chains; culminating in membrane permeabilization and a direct cytotoxic effect on fungal cells. Constitutive expression of pepper (cv. Habanero) LTPs, *CALTPI*, and *CALTPII*, in tobacco conferred enhanced resistance to *Phytophthora nicotianae* and the bacterial pathogen, *Pseudomonas syringae* pv. *tabaci* [23]. When a wheat LTP (*Ltp3F1*) was linked with a chitinase gene from barley (*chi2*) and transformed into carrots, reduced disease symptoms against necrotrophic foliar fungal pathogens, *Alternaria radicola* and *Botrytis cinerea*, were obtained; 95% for *Botrytis* and 90% for *Alternaria* infection compared to 40–50% for single-gene transformants [24].

Phytoalexins are another group of plant secondary metabolites with primarily antifungal activity that have been exploited in the development of fungus-resistant transgenic plants. These metabolites are of low molecular weight (ca. <1,000 Da) and are synthesized or accumulated in response to infection or stress related to infection. Resveratrol is the best studied phytoalexin of grapevine that is synthesized from the precursors, malonyl-CoA and *p*-coumaroyl-CoA, by the action of stilbene synthase. Apparently, the accumulation of resveratrol upon infection triggers the synthesis of a fungal phenol oxidase that results in self-intoxication through the conversion of resveratrol to a more toxic dimer, viniferin. Grapevine stilbene synthase transformation of tomato and papaya [25] introduced increased protection against *Botrytis cinerea* and *Phytophthora palmivora*, respectively. However, in other cases, e.g., strawberry, increased resistance was not obtained [26].

Increased evidence in recent years indicate the possibility of achieving improved resistance in transgenic plants expressing transcription factors that participate

in the regulation of plant defense responses. Five major families of plant transcription factors have been described; basic region/leucine zipper motif (bZIP), zinc finger motif WRKY, MYB, ethylene-responsive element binding proteins (EREBP), and homeodomain proteins. Transcription factors play a crucial role in the transmission of pathogen-derived defense signals. They are involved either in the activation or suppression of downstream defense gene expression or the regulation of cross-talk between different signaling pathways. Presumably the mode of action is by binding promoter elements (W-box) of SAR gene promoters. Thus, the strategy is particularly attractive as the constitutive expression of a transcription factor may also modify the expression of all genes under its regulation. Solano et al. [27] engineered transgenic *Arabidopsis* plants with the ethylene-response-factor 1 (*ERF1*), an early ethylene-response gene that is regulated by the ethylene-insensitive gene EIN3 and, which in turn, regulates the expression of several pathogen responsive genes. The transgenic plants showed a clear reduction in *Botrytis cinerea* infection, with 40–70% of plants remaining symptomless after challenge and never developing macroscopic or microscopic necrosis. *ERF1* also mediated *Arabidopsis* resistance against the soil borne fungi, *Fusarium oxysporum* sp. *conglutinans* and *F. oxysporum* f. sp. *lycopersici* [28]. An ERF transcription factor (*GbERF2*) gene derived from cotton enhanced the resistance of transgenic tobacco to fungal infection by *Alternaria longipes* [29]. The accumulation of *GbERF2* transcripts along with transcripts of PR-proteins, such as *PR-1b*, *PR2*, and *PR4*, were detected.

Rather than using strategies based on the overexpression of PR-proteins or transcription factors that are induced during SAR for enhanced fungal resistance, other researchers have examined the effects of the constitutive expression of signaling molecules that play an important role in the induction of defense proteins. Components of the elicitor-receptor system were used to generate these transgenic plants. The system, explained by the gene-of gene-model, is the major line of plant defense against pathogens. Essentially, a pathogen protein encoded by an avirulence gene (*Avr*) is recognized by a plant protein encoded by a resistance gene (*R*), resulting in the activation of a number of defense mechanisms including the

hypersensitive reaction and the restriction of the pathogen at the site of infection. Thus resistance against fungal pathogens carrying a particular *Avr* gene is introduced via the transfer of the corresponding *R* gene from a resistant plant to a susceptible plant. The success and durability of the strategy correlates strongly on the ability of the *R* gene product to recognize *Avr* proteins secreted by most if not all of the races of the pathogen. During the last decade, several sequences of *R* genes have been characterized into five classes according to their functional domains. Three classes contain leucine-rich repeats. The most common class encodes proteins with an amino-terminal nucleotide-binding site (NBS) and a carboxy-terminal leucine-rich repeat (LRR) and confers resistance against nematodes, sucking insects, viruses, bacteria, oomycetes, and fungi. The proof of concept has been conducted with the model system, tobacco, as well as various economic crops. Rice plants transgenic for the *R* gene alleles, *Pi9*, *Pi2*, and *Piz*, against the rice blast fungus, *Magnaporthe grisea*, were reported highly resistant to 43 isolates collected from 13 countries [30]. In flax, three alleles of the flax rust resistance gene, *L* (*L2*, *L6*, and *L10*) when transformed into a highly susceptible rust cultivar, unequivocally demonstrated resistance against strains of pathogen (*Melampsora lini*) with the corresponding *Avr* gene [31]. Increased resistance was also obtained when transgenic tobacco transformed with the same three flax genes was challenged with two pathogens of tobacco. Increased resistance was conferred against *Cercospora nicotianae* and the oomycete, *Phytophthora parasitica* pv. *nicotianae* in *L6* transgenics and one *L10* transgenic plant [32]. Resistance was attributed to the constitutive expression of defense genes rather than recognition of the pathogens by the *L6* gene product. Of note, the *L6* transgenic plants exhibited a stunted phenotype.

Another arm of the plants' innate defense that is being investigated is that of RNA silencing. Accumulating evidence suggests the involvement of small interfering RNAs (siRNAs) and microRNAs (miRNAs) in the response to fungal pathogens. The involvement of plant siRNAs, short, noncoding RNAs between 20 and 24 nucleotides (nt) in length, have been implicated in the defense against viruses, and more recently bacterial pathogens, and is discussed in a later section of the chapter. miRNAs comprise the 21-nt class of siRNAs

and short-interfering RNAs comprise the 24- nt class of siRNAs. Work with pine and the canker- forming rust fungus, *Cronartium quercuum* f. sp. *fusiforme*, identified 82 putative miRNA targets, including NBS-LRR proteins, receptor-like kinases, laccases, and MYB transcription factors, which are most likely associated with disease responses that restrict fungal growth [33]. Ellendorff et al. [34] showed that RNA silencing plays a role in the defense against *Verticillium* in *Arabidopsis*, but the siRNA involved were not identified. Further research into the targets and processes will invariably have broad implications in potentiating basal defenses against fungal phytopathogens.

Other strategies have explored the development of enhanced resistance through the expression of gene products that target components of the arsenal of the fungus. One approach employed the transformation a susceptible cultivar with sequences encoding polygalacturonase inhibiting proteins (PGIPs). PGIPs, proteins structurally related to several resistance gene products, are expressed in the cell wall of a number of plants and belong to a superfamily of LRR proteins, suggesting involvement in pathogen recognition. These proteins inhibit the activity of fungal endo-polygalacturonases presumably by binding at either the substrate binding site or the underside of the endo-polygalacturonase. The resulting oligogalacturonides induce a range of defense responses. *Arabidopsis* plants overexpressing *PGIP* isolated from pear showed decreases in disease symptoms produced by *Botrytis cinerea* [35]. Similarly, transgenic wheat accumulating bean PGIP developed smaller lesions when challenged with the necrotropic fungus, *Bipolaris sorokiniana* [36]. In earlier studies with transgenic tomato carrying a PGIP transgene isolated from bean, enhanced resistance against *Fusarium oxysporum* f. sp. *lycopersici*, *Botrytis cinerea*, and *Alternaria solani* was not obtained [37].

Another component of the arsenal that some fungi (e.g., *Sclerotinia sclerotiorum*) use to infect plants that is regarded as a pathogenicity factor is oxalate. Oxalic acid apparently assists in initiating infection through acidification, which facilitates cell wall-degrading enzyme activity, through pH-mediated tissue damage, or via the sequestration of calcium ions. Evidence for the utility of hydrogen peroxide-generating enzymes in protective plant responses was provided by the expression of oxalate oxidase, which catalyzes the degradation

of oxalic acid to produce carbon dioxide and hydrogen peroxide (H_2O_2). Essentially increased H_2O_2 production limits the growth of the invading pathogen in planta and triggers the activation of a range of defense responses. Sunflower and tomato transformed with a wheat oxalate oxidase gene exhibited increased resistance to *Sclerotinia sclerotiorum* [38, 39]. Potato (cv. Bintje) and poplar expressing the same gene showed increased resistance to *S. musiva* [40] and the oomycete, *Phytophthora infestans* [41], respectively.

Transgenic Resistance to Bacterial Pathogens

Knowledge of bacteria–host interactions, in particular of the mechanisms of pathogenesis, and efforts to bolster plant defense mechanisms, has provided a strong basis for implementing transgenic approaches for bacterial disease management. Several strategies aiming at the lysis or prevention of bacterial growth have been investigated based on the use of bacterial-resistance genes, avirulence genes, expression of antibacterial peptides that act as bactericidal or bacteriolytic agents, and the application of the concept of pathogen-derived resistance. The latter describes the use of genetic elements from a pathogen's own genome to confer resistance in an otherwise susceptible host via genetic engineering [42].

Members of the NBS-LRR class of disease *R* genes have been employed to provide effective protection against bacteria when transferred into new species. Transformation of a susceptible rice cultivar with the resistance gene *Xa21* that was isolated from a resistant rice line and characterized by positional sequencing and functional genetics conferred resistance to *Xanthomonas oryzae* [43]. Field tests of *Xa21* transgenic rice have shown satisfactory results in the Philippines, India, and the People's Republic of China [44], but deregulation of transgenic *Xa21* rice is still pending. It should be noted that hybrid rice containing *Xa21* was developed by conventional breeding [45]. Similarly, *Xa1* provided resistance to *Xanthomonas oryzae* pv. *oryzae* in transgenic rice [46]. Rice expressing an *R*-gene from maize were resistant against *Xanthomonas oryzae* pv. *oryzicola* [47]. Also, *Bs2* of pepper confers resistance to *Xanthomonas campestris* pv. *vesicatoria* in transgenic tomato [48]. Tomato plants expressing *Pto* showed resistance to *Pseudomonas syringae*

pv. *tomato* and *Xanthomonas campestris* pv. *versicatoria* [49]. The transgenic approach based on *R*-genes circumvents tedious crosses and backcrosses of conventional breeding as well as linkages to undesired traits.

Another approach to confer resistance to bacterial diseases is based on the use of lytic peptides from insects that form pores in bacterial membranes. For example, cecropins are bactericidal peptides from the giant silk moth, *Hyalophora cecropia*, which interact with bacterial membranes and form transient ion channels. The peptides are active against a wide range of plant pathogenic bacteria including *Erwinia carotovora*, *E. amylovora*, *Pseudomonas syringae*, *Ralstonia solanacearum*, and *Xanthomonas campestris*. Native, mutant (SB37 and MB39) or synthetic polypeptides (Shiva-1 and D4E1) that are designed for increased activity have been used for engineered resistance against bacterial diseases. Infection caused by *Erwinia carotovora* sp. *atroseptica* is reduced in transgenic potato expressing the Shiva-1 and SB-37 lytic peptide analogs [50]. Transgenic apple trees expressing SB-37 showed increased resistance to *E. amylovora* in field tests [51]. Poplar expressing D4E1 had more resistance to *Agrobacterium tumefaciens* and *Xanthomonas populi* than control nontransformed trees [52, 53]. Attacin is another small protein from the giant silk moth with antibacterial activity that inhibits synthesis of the outer bacterial membrane. Transgenic potato attacin expression has enhanced resistance to *Erwinia carotovora* sp. *atroseptica* [50]. Attacin [54] as well as lactoferrin [55] provide enhanced resistance against *Erwinia amylovora* in pear. Similarly, transgenic apple expression of the attacin E gene exhibit enhanced resistance to *Erwinia amylovora* [56] even under field conditions [57].

Lysozymes from various sources have been exploited in the development of transgenic resistance against bacterial diseases. Lysozymes are ubiquitous bacteriolytic enzymes that cleave the murein layer of bacterial peptidoglycan, resulting in a weakening of the bacterial cell wall. A chimeric bacteriophage T4 lysozyme fused to a plant signal peptide reduces the susceptibility of transgenic potato plants to *Erwinia carotovora* spp. *atroseptica* [58]. Transgenic apple trees with the bacteriophage T4 lysozyme exhibited significant resistance to fire blight infection [59] and the bovine lysozyme isoform was shown to confer

resistance to *Xanthomonas campestris* in tomato, rice, and potato [60].

Tachyplesin, an antibacterial peptide isolated from the horseshoe crab (*Tachyplesus tridentatus*), reduces the incidence of rubber rot caused by *E. carotovora* in transgenic potato [61].

Reactive oxygen species play important roles in various defense responses of plants, including pathogen infection. A prolonged local oxidative burst is one of the earliest events correlated with plant resistance at the site of pathogen invasion; see the previous section for detailed information on the role of oxalate oxidase in plant defense responses. Transgenic potato tubers expressing a glucose oxidase from *Aspergillus niger* for the production of large amounts of hydrogen peroxidase exhibit strong resistance to *Erwinia carotovora* [62].

Many bacteria control virulence factor expression by a cell-to-cell communication system referred to as quorum sensing. Disruption of bacterial quorum sensing has been proposed as a disease management strategy and several techniques with the potential to disrupt quorum sensing have been investigated. The enzyme *N*-acyl homoserine lactone (AHL)-lactonase (AiiA) isolated from the soil-borne *Bacillus* sp. 240B1, which catalyzes the degradation of AHL and attenuates symptoms caused by plant pathogens by degrading the quorum-sensing autoinducer AHL of the soft rot pathogen, *Erwinia carotovora*, induces enhanced resistance in transgenic potato [63].

Several pathovars of *Pseudomonas syringae* produce extracellular toxins that often increase their virulence toward plants. Phaseolotoxin is one such example. Phaseolotoxin is a modified tripeptide that inhibits the plant enzyme, ornithine carbamoyl transferase (OCTase), and is produced by the *P. syringae* pvs. *phaseolicola* and *actinidiae* as well as by a single strain of *P. syringae* pv. *syringae*, CFBP3388. Transgenic bean expressing *argK* that codes for a phaseolotoxin-insensitive OCTase show resistance to *Pseudomonas syringae* pv. *phaseolicola* [64].

An example of pathogen-derived resistance against a bacterial disease, crown gall, is reported by Krastanova et al. [65]. Based on an earlier report on the use of a truncated virE2 gene construct lacking the 215 C-terminal amino acids to confer resistance to grown gall disease in tobacco [66], these authors used

a truncated form of *virD2* from *Agrobacterium tumefaciens* strains C58 and A6, and *Agrobacterium vitis* strain CG450 to engineer resistance to crown gall disease in transgenic grape rootstocks. Resistant lines show a substantial reduction in tumorigenicity and develop very small sized galls [65].

Other researchers used a different approach against the soil-borne bacterium, *Agrobacterium tumefaciens*, that utilized oncogenes of the phytopathogen. *Agrobacterium tumefaciens* possesses several oncogenes (e.g., *iaaM* and *ipt*) that trigger de novo synthesis of auxins and cytokinins to generate tumors following horizontal gene transfer into the plant genome. By expressing two self-complementary RNA constructs designed to initiate RNA interference (RNAi) of the *iaaM* and *ipt* oncogenes of *A. tumefaciens*, resistance to crown gall disease development was achieved in transgenic tomato plants. Transformed tomato lines display between 0.0% and 24.2% tumorigenesis, whereas controls averaged 100% tumorigenesis following stem inoculation with various strains of *A. tumefaciens* [67]. This mechanism of resistance is based on RNA silencing rather than the highly specific receptor–ligand-binding interactions characteristic of traditional plant resistance genes.

Manipulation of the plant's innate defense signaling pathways offers an alternative approach to conferring resistance against bacterial diseases by controlling a large number of induced genes either directly or indirectly. The nonexpressor of PR genes, *NPR1*, is a key regulatory gene of the salicylic acid-mediated systemic acquired resistance in *Arabidopsis thaliana*. Moreover, *NPR1* plays a key role in ethylene and jasmonic acid signaling pathways that are involved in induced systemic resistance. Plants expressing *AtNPR1* are resistant to bacterial, fungal, and viral pathogens [68–70]. Similarly, transgenic carrots expressing *AtNPR1* are resistant to *Xanthomonas hortorum* as well as to fungal pathogens [71] and transgenic tomato plants expressing *AtNPR1* show enhanced resistance to *Ralstonia solanacearum* and to a lesser extent to *Xanthomonas campestris* [68]. When introduced into rice, *AtNPR1* confers resistance to the bacterial pathogen, *Xanthomonas oryzae* pv. *oryzae* [72]. Overexpression of the rice *AtNPR1* homolog, *OsNPR1*, also enhances resistance to *X. oryzae* pv. *oryzae* in rice [73] and overexpression of an apple

NPRI confers increased resistance to *Erwinia amylovora* in transgenic apple cultivar 'Galaxy' and transgenic apple rootstock M26 [74].

The involvement of endogenous siRNA and miRNAs in mediating the triggering of defense mechanisms against bacteria is an area of considerable interest. In spite of tremendous progress at identifying and sequencing siRNAs and miRNAs, information available on the involvement of endogenous small RNAs at mediating defense regulation is scarce. One miRNA (miR393) was recently shown to contribute to basal defense against bacteria by targeting auxin-signaling components [75]. Agorio and Vera [76] characterized an *Arabidopsis* ocp (overexpressor of cationic peroxidase) mutant that overexpresses the H₂O₂-responsive Ep5C promoter. The ocp11 mutant exhibits enhanced disease susceptibility to several virulent and avirulent *Pseudomonas syringae* strains. OCP11 was cloned and found to encode ARGONAUTE4 (AGO4), a component of the pathway that mediates the transcriptional gene silencing associated with siRNA. The mutant allele, ago4-1, was examined and found to be compromised in resistance to *P. syringae* [76]. This work provided insight into the involvement of small RNA gene silencing pathways in resistance to bacterial pathogens. Silencing the expression of key genes involved in pathogen–host interactions and disease susceptibility is opening new avenues for bacterial disease management.

Transgenic Resistance to Viral Pathogens

Transgenic resistance generally complements conventional breeding methodologies for the development of virus-resistant crops; however, the approach devised to engineer resistance against viral diseases differs from strategies employed in engineering resistance against fungal and bacterial diseases. This is due to the fact that viruses utilize cellular organelles and metabolic pathways for their replication. As such, they interfere with the basic functions in a living cell, unlike fungi and bacteria, which produce specialized reproductive structures or reproduce by binary fission. The approach to engineering resistance to viral diseases in crops essentially consists of activating anti-viral pathways of a natural, innate, and potent defense mechanism against viruses called RNA silencing [77–82].

Prior to the discovery of RNA silencing in the late 1990s and early 2000s, the concept of PDR [42] was commonly applied to engineering resistance against viral diseases. Initially, the viral coat protein gene was the preferred construct used to confer resistance to viruses in plants. However, it soon became apparent that almost any sequence derived from a viral genome, i.e., coat protein gene, movement protein gene, proteinase gene, RNA-dependent RNA polymerase gene, 5' and 3' noncoding regions, satellite RNA, defective interfering RNA, could provide resistance. Validated first in tobacco plants expressing the coat protein gene of *Tobacco mosaic virus* (TMV) [83], PDR was subsequently applied to horticultural crops for engineered resistance against viral diseases. In the first field trial of transgenic plants engineered for virus resistance, tomato plants expressing the coat protein gene of TMV were evaluated for resistance to mechanical inoculation by TMV [84]. Only 5% of the transgenic plants were symptomatic at the end of the trial compared with 99% of the nontransformed control plants. This study indicated that overexpression of the coat protein gene provided practical control of TMV under field conditions. Also, inoculated transgenic and uninoculated nontransformed plants had identical fruit yield, indicating that the transformation process and expression of the TMV coat protein gene did not alter the horticultural performance of the transgenic tomato plants.

The efficiency of viral genes at conferring resistance against vector-mediated virus transmission was first shown with cucumber plants engineered for resistance to *Cucumber mosaic virus* (CMV). Plants expressing the coat protein gene of CMV showed a significantly reduced incidence of CMV and a lower percentage of symptomatic plants than nontransformed control plants following CMV inoculation via aphid vectors [85]. In these studies, mechanically inoculated cucumber plants dispersed throughout the field provided reliable sources of inoculum for natural aphid populations to vector CMV. This approach coupled with the fact that field trials were established at a time of abundant endemic aphid flights caused sufficient disease pressure to make inferences about disease progress, resistance, and yield [85]. Subsequently, many other studies have confirmed the usefulness of engineered resistance at providing practical control of aphid-transmitted virus diseases [6].

PDR offers unique solutions to infection by multiple viruses, for example, by co-engineering and co-transferring genes from several viruses into a single host plant. The usefulness of multiple viral genes to control mixed virus infections was demonstrated early on with potato plants expressing the coat protein genes of *Potato virus X* (PVX) and *Potato virus Y* (PVY) [86, 87]. Potato line 303 was highly resistant to infections by PVX and PVY in the field [87]. Later, summer squash plants expressing coat protein gene constructs of CMV, *Zucchini yellow mosaic virus* (ZYMV) and/or *Watermelon mosaic virus* (WMV) were engineered for resistance to single viruses or combinations of these three viruses [88]. Among summer squash engineered for multiple virus resistance, line ZW-20 expressing the coat protein genes of ZYMV and WMV was highly resistant, regardless of whether the infections were initiated by mechanical inoculations or mediated by aphid vectors [88, 89]. Similarly, transgenic line CZW-3 expressing the coat protein genes of CMV, ZYMV, and WMV was highly resistant to mixed infections by these three viruses following aphid transmission [88, 90].

Following the initial discovery by Powell Abel et al. [83], the viral coat protein gene from various viruses was introduced into numerous economically important crop species with the aim of achieving resistance. It was initially believed that resistance was provided by the viral protein itself via a mechanism involving excess plant-expressed coat protein that interfered with the uncoating step in viral replication [91]. However, it soon became apparent that resistance could be achieved in transgenic plants producing low or undetectable levels of coat protein [92]. The mechanism underlying the engineered resistance involved degradation of the transgene-derived mRNA into small fragments in a sequence-specific manner. This phenomenon was subsequently referred to as RNA silencing [78–82]. This antiviral plant defense mechanism is initiated by double stranded RNA (dsRNA) structures that are identical to the RNA to be degraded [93]. Silencing is associated with the production of 21–24 nt duplexes called small interfering RNAs (siRNAs) [94, 95]. The siRNAs are produced from dsRNA precursors by an endonuclease known as Dicer and become incorporated and converted to single stranded RNAs (ssRNAs) in an Argonaute-containing-

RNA-induced silencing complex (RISC) that targets RNA for cleavage [80, 82, 96, 97].

RNA silencing is an innate and potent plant response to virus infection and a natural example of the concept of PDR that has provided new and unprecedented insights into virus–host interactions. Several approaches have been used to express dsRNA cognate to viral RNA for activation of RNA silencing. Expressing sense and antisense viral genes or inverted repeat viral genes to express hairpin RNAs (hpRNA) for the formation of duplex RNA are some of the most recent strategies used to engineer resistance against viruses [81]. For example, intron-spliced hairpin RNA (ihpRNA), ihpRNA overhang, and ihpRNA spacer were evaluated for resistance to PVY [98, 99]. The ihpRNA was found to be the most efficient construct to conferring resistance to PVY with 90% of the plants exhibiting RNA silencing [99]. The same strategy based on the use of highly conserved genetic segments of several viruses into a single transgene construct achieved multiple virus resistance [100].

Artificial plant micro RNAs (amiRNAs) have been used also to engineer virus resistance in plants. The *Arabidopsis thaliana* pre-miR159a precursor was used to generate two amiRNAs¹⁵⁹ (amiR-P69¹⁵⁹ and amiR-HC-Pro¹⁵⁹) with sequences complementary to *Turnip yellow mosaic virus* (TYMV) and *Turnip mosaic virus* (TuMV), respectively [101]. The amiR-P69¹⁵⁹ was designed to target the viral suppressor P69 of TYMV while amiR-HC-Pro¹⁵⁹ targeted the viral suppressor HC-Pro of TuMV. Transgenic plants carrying both transgenes expressed the corresponding amiRNAs and showed specific resistance to TYMV and TuMV. Low temperatures had no substantial effect on miRNA accumulation [101]. Similarly, the miR171 of *Nicotiana benthamiana* was used to target the 2b gene of CMV and confer resistance to CMV [102]. Although very promising, amiRNAs have not been used yet in crops for virus resistance.

Strategies other than PDR and RNA silencing have also been applied to confer resistance to viral diseases in plants. Peptides or protein microdomains designed to block protein-binding sites have been used. For example, peptide aptamers designed to microdomains of the nucleoprotein of *Tomato spotted wilt virus* (TSWV) that interact with one of the N-protein homomultimerization domains confer resistance to TSWV and

other tospoviruses in *Nicotiana* species [103]. Similarly, antibody-based resistance has been investigated as an innovative strategy to confer virus resistance in plants. Hybridoma-derived single-chain variable antibody fragments (scFv) were engineered for resistance to various viruses in model plants [104–108]. A number of plant proteins with antiviral activity have been identified and exploited for engineered resistance to viruses in plants. A class of proteins termed ribosome-inactivating proteins [109], protease inhibitors (cystatins) [110], and the interferon-regulated 2–5A system [111, 112] were used successfully to confer virus resistance.

Status of Commercialized Disease Resistant Transgenic Crops

Since the early studies of Powell Abel and coworkers [83], numerous cereal, vegetable, legume, flower, forage, turf, and fruit crops expressing virus-derived gene constructs have been created [6]. Many have been tested under field conditions and shown to be highly resistant to virus infections. Among the transgenic crops produced and evaluated in the field, two lines of transgenic summer squash, i.e., line ZW-20 resistant to ZYMV and WMV, and line CZW-3 resistant to ZYMV, WMV, and CMV [88], and papaya resistant to PRSV [113] have been deregulated and released for commercial use in the USA. Summer squash line ZW-20 received exemption status in 1994 and was the first disease-resistant transgenic crop to be commercialized. Summer squash line CZW-3 was deregulated and commercialized in 1996. Five transgenic zucchini cultivars and six transgenic straightneck or crookneck yellow squash cultivars derived from lines ZW-20 and CZW-3 were developed by conventional breeding. Their adoption is increasing steadily since their initial release in 1996. In 2006, the adoption rate was estimated to 22% (3,250 ha) across the country with an average rate of 70% in New Jersey and 20% in Florida, Georgia, and South Carolina [114]. Papaya expressing the coat protein gene of PRSV was deregulated and commercialized in Hawaii in 1998. PRSV is a major limiting factor to papaya production in Hawaii and around the world. After extensive testing, PRSV-resistant papaya was released in 1998 as devastation caused by the virus reached record

proportions in the archipelago's main production region [113]. The impact of PRSV-resistant papaya on the papaya industry in Hawaii is evidenced by its rapid adoption rate. In 2000, the first wave of transgenic papaya bore fruit on more than 42% of the total acreage, and by 2009, transgenic papaya cultivars were planted on more than 90% of the total papaya land in Hawaii (780 of 866 total hectares) ([114], D. Gonsalves, personal communication October 2009).

Tomato and sweet pepper resistant to CMV and papaya resistant to PRSV are also released in the People's Republic of China [115, 116]. Limited, if any, information is available on their adoption rate. Two virus-resistant potato lines were deregulated also in 1998 and 2000 in the USA. After failed attempts to create a potato line resistant to *Potato leafroll virus* (PLRV) by coat protein gene expression, lines expressing a PLRV replicase gene were created, field tested, deregulated, and commercialized [117]. Resistance to PLRV was stacked with the coat protein gene of PVY. Many growers in the Pacific Northwest, Midwest USA and Canada are growing virus-resistant transgenic potato, and a breakdown in resistance has not been reported, neither any detrimental impact on the environmental or human health. Nonetheless, virus-resistant potato was withdrawn from the market after the 2001 growing season due to the reluctance of large processors and exporters to adopt these products [117].

Although not released yet, the transgenic plum cultivar 'Honeysweet' resistant to *Plum pox virus* (PPV) and another PRSV-resistant papaya are under consideration for deregulation in the USA. Plum trees expressing a coat protein gene of PPV are highly resistant to PPV infection [118–121]. The US Department of Agriculture (USDA) Animal and Plant Health Inspection Service (APHIS) granted this plum cultivar deregulated status [122] and the Food and Drug Administration (FDA) has deemed a premarket review of the 'Honeysweet' unnecessary. Presently, the Environmental Protection Agency (EPA) is examining deregulation petitions for 'Honeysweet'. Papaya line X17-2 is also being considered for deregulation in the United States. Line X17-2 differs from the previously deregulated Hawaiian papaya in that it expresses the coat protein gene of a Florida isolate of PRSV and is suitable for cultivation in Florida [123]. APHIS and

FDA granted X17-2-deregulated status and EPA is presently reviewing a deregulation petition [124].

Despite considerable progress in identifying genes that confer resistance against fungal and bacterial pathogens and testing their efficacy in transgenic plants under greenhouse conditions, it appears that the challenge is to translate the results to an effect in the field under continuous pathogen pressure. A search of the field testing release permit database in the USA reveals over 25 applications for field testing fungus-resistant transgenic soybean, corn, grape, strawberry, tobacco, lettuce, wheat, and American elm between 1993 and 2009. Both single traits such as resistance to Dutch Elm disease, ear mold, or downy mildew, were assessed as well as multiple traits against *Botrytis cinerea*, bacterial pathogens, nematodes, abiotic tolerance, glyphosate tolerance, and altered nutritional properties (oil and seed composition). Based on the FAO Biotechnology in Developing Countries Database (FAO-BioDeC), field trials have been conducted in South Africa with transgenic strawberry expressing a phytoalexin synthesis gene and transgenic cotton with resistance to *Verticillium* and *Fusarium* in China. A few countries in Latin America, namely, Argentina, Brazil, and Cuba, are working on transgenic fungal resistance in tropical fruit trees, with some results already in the field. Apart from small projects in China, Thailand, and Pakistan addressing potato and wheat wilt, tomato wilt, and blight in rice, far fewer initiatives have been devoted to the development of transgenic cultivars with resistance against bacterial diseases.

To date, no transgenic cultivar with fungal or bacterial resistance has been released into commerce. The outcomes of field experiments have been variable and not in all cases is the resistance observed in the greenhouse extended to the field [125, 126]. In general, the constitutive expression of antimicrobial proteins in the transgenic host appears to confer partial resistance against these pathogens or increased tolerance at best, rather than absolute resistance. It is also clear that the resistance is limited to a few pathogens. The findings of Punja and Raharjo [127] further suggest that the levels of resistance in transgenic plants are influenced by interactions between endogenous and transgene defense compounds. These authors showed that transgenic carrots carrying a chitinase gene from tobacco exhibited enhanced resistance against *Botrytis cinerea*,

Rhizoctonia solani, and *Sclerotium rolfsii*. However, the transgenic plants reacted similarly to *Alternaria radicini* and *Thielaviopsis basicola* as nontransgenic carrots. Further, improved resistance was not obtained against the same sleuth of pathogens and transgenic carrots expressing a chitinase gene from petunia. Indistinguishable disease reactions to *A. cucumerina*, *B. cinera*, *Collectotrichum lagenarium*, and *R. solani* were observed with transgenic and nontransgenic cucumber plants transformed with the chitinase gene from three sources, tobacco and petunia, as above, as well as bean.

Conclusions

Plant diseases caused by fungi, bacteria, and viruses are responsible for enormous losses in cultivated crops worldwide. Disease management relies on prevention, cultural practices, biological control, use of chemical pesticides and insecticides, antibiotic sprays, and selection of host resistant genotypes. These approaches have had limited success in the prevention or cure of diseases. Moreover, the frequent use of biocides, especially in subtherapeutic doses, is leading to the rapid development of resistance in certain fungi and bacteria. In addition, preventing fungal infection not only helps in combating yield losses but also keeps crops free of toxic compounds that are produced by some pathogenic fungi and referred to as mycotoxins. These compounds can affect the immune system and disrupt hormone balances or can be carcinogenic. Therefore, there is an urgent need to develop alternative ways to control infections caused by fungal, bacterial, and viral pathogens in agriculture.

The incorporation of specific disease-resistant traits in plants through genetic engineering offers a means to prevent disease-associated losses. For viruses, the concept of PDR [42] provided unique opportunities for innovative solutions to control virus diseases by developing resistant crops expressing genetic elements derived from a virus's own genome. Thirty years after the inception of PDR, various transgenic crops have been engineered for virus resistance and virus-resistant papaya, summer squash, sweet pepper, and tomato have been deregulated and released commercially. Applying the concept of PDR and the antiviral pathways of RNA silencing provides unique opportunities for developing virus-resistant crops and implementing

efficient and environmentally sound management approaches to mitigate the impact of viral diseases. Based on the tremendous progress, the prospects of further advancing this innovative technology for practical control of virus diseases are very promising.

Engineered resistance to fungal and bacterial diseases is more complex than engineered resistance to viral diseases. This is due to the complex nature of fungi and bacteria, as well as complex interactions with their hosts. Therefore, in spite of remarkable progress and promising reports from greenhouse evaluations and field tests, no cultivar with engineered resistance to fungi and bacteria have been deregulated yet or commercially released. The fact that transgenic crops with resistance to fungi or bacteria might not have provided equivalent levels of plant protection compared to conventional management strategies in practice could account for the reluctance to deliver transgenic plans or seeds to growers. Also, the degree of resistance to fungal and bacterial diseases achieved so far might be of limited practical use for disease management. However, new insights into pathogen–host interactions and the application of siRNA and microRNA technologies will undoubtedly identify new targets and open new horizons for the management of fungal and bacterial diseases.

Future Directions

Transgenic virus resistance is undoubtedly the most advanced of the applications of biotechnology for the management of plant pathogens. There is the documented safe release of virus-resistant transgenic crops over the past 14 years in the USA, in addition to a significant amount of evidence that these crops have little to no detrimental impact on animal health and the environment beyond those of conventional agricultural crops [6]. Given the discovery and elucidation of the antiviral pathways of RNA silencing, various new approaches have been used to develop transgenes more likely to stimulate RNA silencing via the design of transgenes that will form dsRNA structures in planta. These approaches seemingly have an advantage over a full-length coat protein gene in the sense orientation as they are generally unable to produce a functional protein, thus alleviating concerns arising from the presence of the coat protein in plant material. Some

nonviral sources of virus resistance, such as host resistance genes and the silencing of host genes that are necessary for viral replication [81] have also been investigated and could theoretically alleviate concerns about synergism, recombination, and transencapsidation. Work with transcription factors, in particular the ERF subfamily transcription, provided evidence of increased resistance to bacterial, fungal, and viral pathogens as well as abiotic stress via overexpression of the soybean *GmERF3* transcription factor in tobacco [128]. The mechanism of the effect remains unknown; it is likely that a range of effects such as elevated antioxidant capacity, induced expression of chitinases, or antimicrobial compounds in *GmERF3* transgenic plants and the suppression of viral spread as a consequence of the reduced virus propagation are responsible. The commercial potential of these technologies remain largely undiscussed and untested.

On the other hand, engineering resistance to fungal and bacterial diseases has proven more recalcitrant. Whereas crops engineered for resistance against viral pathogens are in the market, no commercial transgenic product with enhanced resistance against fungal or bacterial diseases is currently available. These organisms have developed numerous survival strategies and more complex interactions with the plant host. As a result, it has proven more difficult to engineer resistance against these pathogens; far beyond designing a simple defense strategy involving the use of a single gene.

Strides in deciphering the siRNAs that are induced or repressed in response to pathogen attack will invariably contribute to new technologies for the development of transgenic resistance against fungal and bacterial pathogens. Microarray analysis with more wild-type plants and RNA-silencing mutants should reveal the relationship between the physiological processes and the associated disease-resistant phenotype. In the interim, the use of stacks or pyramids of *R* genes either from self or non-host plant species is perhaps the best approach available for the development of transgenic resistance against these pathogens. Presumably, this strategy would not affect resistance to nontarget pathogens while providing durable and broad-spectrum resistance and not interfering with the tight controls of the plants' defense systems. But the full potential of *R* gene expression to provide wide range resistance cannot be realized as long as its use is

clouded by controversy over what it does and how well it works. The major limitation of the use of *R* genes in the development of transgenic resistance lies in the durability of the resistance, which is driven by the highly specific recognition of the elicitor molecule of the pathogen. Pathogens are usually able to overcome *R* gene-mediated recognition through the accumulation of mutations or through the loss of the elicitor gene [129, 130]. Moreover, not in all cases is resistance obtained after transfer to species that are not closely related [32]. Although a plethora of data is available on the identity of genes and gene products involved in plant resistance, knowledge of pathogenesis is relatively primitive. Continued sequencing of crop plant genomes as well as pathogen genome analysis will invariably lead to decoding of the functions of various defense genes as well as pathways and the isolation of more *R* genes. With this knowledge, the precise manipulation of *R* genes to permit binding to proteins of pathogens and the activation of defense responses will be possible, and facilitate the design of synthetic *R* genes with desired specificities. For example, work with 13 alleles (*L*, *L1* to *L11*, and *LH*) from the *L* locus from flax suggests a role for the TIR region and LRR region of *L* proteins in the flax–flax rust interaction [131]. Most of the sequence variation between the three resistance genes, *Pi9*, *Pi2*, and *Piz*, is confined to the LRR domains, indicating that this domain plays a major role in determination of *Magnaporthe grisea* resistance specificities [30]. Further, Jia et al. [132] provided evidence that a single amino acid change in the *Pi-ta* LRD (leucine rich domain) or the *AVR-Pita*₁₇₆ protease motif disrupted interaction between the two and subsequent a loss of resistance against *M. grisea*, the etiological agent of rice blast disease. As more data become available on the structural information and on the interactions between resistance proteins and pathogen ligands, the tweaking of *R* gene domains to accommodate resistance of pathogens infecting distantly related crop species will be likely possible. Moreover, this could be extended to antimicrobial proteins that target components of the pathogen's arsenal, e.g., PGIP. Manipulation of their primary structure and hence determinants of specificity could theoretically generate novel recognition specificities.

Multiple races and range in plant hosts are major issues facing the development of durable disease

resistance against fungal and bacterial pathogens. Moreover, in some instances, the expression of enhanced resistance is accompanied by reduced growth or altered morphology or development, given that the plant has been reprogrammed into defense mode. Precise control of transgene expression is thus pivotal; restricted expression at the site of infection and quick induction is crucial as well as a quick response to a wide variety of pathogens [133]. One option is to use pathogen-inducible promoters. Given that the available pathogen-inducible promoters show some patterns of background expression, transcriptomics will play an important role in the identification of more suitable promoters. Work with *Arabidopsis* has already identified potential candidates [134], as have investigations into synthetic promoters combining *cis* elements (W box and GCC-like box) that are conserved across species for restricted expression exclusively at the site of attempted pathogen invasion [135]. Other approaches are considering defensin promoters [136], while others have utilized tissue-specific promoters with some success [137]. The endogenous endothionin *Thi2.1* gene in *Arabidopsis thaliana* gene in tomato under the control of a fruit-inactive promoter (RB7) provided significant levels of enhanced resistance to bacterial and Fusarium wilt [137]. More recently, one study targeted the expression of the PR-protein in leaf trichomes, the preferential port for plant colonization by fungi. The gene encoding an α -1,3-glucanase from *Trichoderma harzianum* was fused to the ATP promoter that confers high expression levels in trichome cells and transferred to *Arabidopsis* [138]. Increased resistance was obtained against *Botrytis cinerea* in needle-wounded and spray assays with transgenic plants. Up to 20% fewer total leaf infections were observed.

In summary, transgenic resistance against viral, fungal, and bacterial pathogens is being addressed on many fronts. The challenge is to develop durable, broad-spectrum resistance, given the diversity of strategies that pathogens deploy and their ability to rapidly adapt. Clearly, an integrated approach based on the knowledge of the defense system of the plant, the arsenal of the pathogen to cause disease and the key genes involved in pathogenesis is required for the design of effective transgenic management strategies. To date, tremendous progress has been made for viruses. Encouraging results have been obtained for

various fungi and bacteria. The prospects of further advancing transgenic crops for practical control of fungal, bacterial, and viral diseases are very promising.

Bibliography

Primary Literature

- Agrios G (2005) Plant pathology, 5th edn. Elsevier/Academic, San Diego, CA
- Herrera-Estrella L, Depicker A, van Montagu M, Schell S (1983) Expression of chimaeric genes transferred into plant cells using a Ti-plasmid-derived vector. *Nature* 303:209–213
- Bevan MW, Flavell RB, Chilton MD (1983) A chimaeric antibiotic resistance gene as a selectable marker for plant cell transformation. *Nature* 304:184–187
- Fraley RT, Rogers SG, Horsch RB, Sanders PR, Flick JS, Adams SP, Bittner ML, Brand LA, Fink CL, Fry JS, Galluppi GR, Goldberg SB, Hoffmann NL, Woo SC (1983) Expression of bacterial genes in plant cells. *Proc Natl Acad Sci USA* 80:4803–4807
- Murai N, Sutton DW, Murray MG, Slightom JL, Merlo DJ, Reichert NA, Sengupta-Gopalan C, Stock CA, Barker RF, Kemp JD, Hall TC (1983) Phaseolin gene from bean is expressed after transfer to sunflower via tumor-inducing plasmid vectors. *Science* 222:476–482
- Fuchs M, Gonsalves D (2007) Safety of virus-resistant transgenic plants two decades after their introduction: lessons from realistic field risk assessment studies. *Annu Rev Phytopathol* 45:173–202
- Lindow SE, Hecht-Poinar EI, Elliot VJ (eds) (2002) Phyllosphere microbiology. American Phytopathological Society, St. Paul, MN
- Matthews R (1973) Induction of disease by viruses, with special reference to *Turnip yellow mosaic virus*. *Annu Rev Phytopathol* 11:147–168
- Struck C (2006) Infection strategies of plant parasitic fungi. In: Cooke BM, Gareth Jones D, Kaye B (eds) *The epidemiology of plant diseases*. Springer, Netherlands, pp 117–137
- Oerke EC (2006) Crop losses to pests. *J Agric Sci* 144:31–43
- Zadoks JC, Schein RD (1979) *Epidemiology and plant disease management*. Oxford Press, New York
- Scorza R (2003) Breeding and molecular genetics. In: Auxt Baugher T, Singha S (eds) *Concise encyclopedia of temperate tree fruit*. The Haworth Press Inc, Binghamton, NY
- Brogli K, Chet I, Holliday M, Cressman R, Biddle P, Knowlton S, Mauvais CJ, Brogli R (1991) Transgenic plants with enhanced resistance to the fungal pathogen *Rhizoctonia solani*. *Science* 254:1194–1197
- Lorito M, Woo SL, Garcia Fernandez I, Colucci G, Harman G, Pintor-Toros J, Filippone E, Muccifora S, Lawrence C, Zoina A, Tuzun S, Scala F (1998) Genes from mycoparasitic fungi as a source for improving plant resistance to fungal pathogens. *Proc Natl Acad Sci USA* 95:12734–12739
- Maziah M, Sariah M, Sreeramanan S (2007) Transgenic banana rastali (AAB) with b 1, 3-glucanase gene for tolerance to fusarium wilt race 1 disease via *Agrobacterium*-mediated transformation system. *Plant Pathol J* 6:271–282
- Bolar JP, Norelli JL, Harman GE, Brown SK, Aldwinckle HS (2001) Synergistic activity of endochitinase and exochitinase from *Trichoderma atroviride* (*T. harzianum*) against the pathogenic fungus (*Venturia inaequalis*) in transgenic apple plants. *Trans Res* 10:533–543
- Anand A, Zhou T, Trick H, Gill B, Bocku W, Muthukrishnan S (2003) Greenhouse and field testing of transgenic wheat plants stably expressing genes for thaumatin-like protein, chitinase and glucanase against *Fusarium graminearum*. *J Exp Bot* 54:1101–1111
- Ó'Brien PA, MacNish GC, Milton NM (2001) Transgenic tobacco plants show different resistance to *Rhizoctonia solani* AG 4 and AG 8. *Aust Plant Pathol* 30:221–225
- Eppl P, Apel K, Bohlmann H (1997) Overexpression of an endogenous thionin enhances resistance of *Arabidopsis* against *Fusarium oxysporum*. *Plant Cell* 9:509–520
- Chan Y, Prasad V, Sanjaya C, Chen K, Liu P, Chan M, Cheng C (2005) Transgenic tomato plants expressing an *Arabidopsis* thionin (*Thi2.1*) driven by fruit-inactive promoter battle against phytopathogenic attack. *Planta* 221:386–393
- Anuradha TS, Divya K, Jami SK, Kirto PB (2008) Transgenic tobacco and peanut plants expressing a mustard defensin show resistance to fungal pathogens. *Plant Cell Rep* 27:1777–1786
- François I, De Bolle M, Dwyer G, Goderis I, Wouters P, Verhaert P, Proost P, Schaaper W, Cammue B, Broekaert W (2002) Transgenic expression in *Arabidopsis* of a polyprotein construct leading to production of two different antimicrobial proteins. *Plant Physiol* 128:1346–1358
- Sarowar S, Kim J, Deok Kim K, Kook Hwang B, Han Ok S, Sheop Shin J (2009) Overexpression of lipid transfer protein (LTP) genes enhances resistance to plant pathogens and LTP functions in long-distance systemic signaling in tobacco. *Plant Cell Rep* 28:419–427
- Jayaraj J, Punja Z (2007) Combined expression of chitinase and lipid transfer protein genes in transgenic carrot plants enhances resistance to foliar fungal pathogens. *Plant Cell Rep* 26:1539–1546
- Zhu YJ, Tang CS, Moore P (2005) Increased disease resistance in papaya by transforming a pathogen-inducible stilbene synthase gene. *Acta Hort* 692:107–114
- Hanhineva K, Kokko H, Rogachev I, Aharoni A, Karenlampi S (2009) Stilbene synthase gene transfer caused alterations in the phenylpropanoid metabolism of transgenic strawberry (*Fragaria ananassa*). *J Exp Bot* 60:2093–2106
- Solano R, Stepanova A, Chao QM, Ecker JR (1998) Nuclear events in ethylene signaling: a transcriptional cascade mediated by ETHYLENE-INSENSITIVE3 and ETHYLENERESPONSE-FACTOR1. *Gene Dev* 12:3703–3714
- Berrol-Lobo M, Molina A (2004) ETHYLENE RESPONSE FACTOR 1 mediates *Arabidopsis* resistance to the soil-borne

- fungus *Fusarium oxysporum*. Mol Plant-Microbe Interact 17:763–770
29. Zuo K, Qin J, Zhao J, Ling H, Zhang L, Cao Y, Tang K (2007) Over-expression *GbERF2* transcription factor in tobacco enhances brown spots disease resistance by activating expression of downstream genes. Gene 391:80–90
 30. Qu S, Liu G, Zhou B, Bellizzi M, Zeng L, Dai L, Han B, Wang GL (2006) The broad-spectrum blast resistance gene *Pi9* encodes a nucleotide binding site-leucine-rich repeat protein and is a member of a multigene family in rice. Genet 172:1901–1914
 31. Ellis J, Dodds P, Lawrence G (2007) Flax rust resistance gene specificity is based on direct resistance-avirulence protein interactions. Annu Rev Phytopathol 45:289–306
 32. Frost D, Way H, Howles P, Luck J, Manners J, Hardham A, Finnegan J, Ellis J (2004) Tobacco transgenic for the flax rust resistance gene *L* expresses allele-specific activation of defense responses. Mol Plant-Microbe Interact 17:224–232
 33. Lu SF, Sun YH, Amerson H, Chiang VL (2007) MicroRNAs in loblolly pine (*Pinus taeda* L.) and their association with fusiform rust gall development. Plant J 51:1077–1098
 34. Ellendorff U, Fradin EF, de Jonge R, Thomma BP (2009) RNA silencing is required for *Arabidopsis* defence against *Verticillium* wilt disease. J Exp Bot 60:591–602
 35. Powell AL, Van Kan J, Ten Have A, Visser J, Greve LC, Bennett AB, Labavitch JM (2000) Transgenic expression of pear PGIP in tomato limits fungal colonization. Mol Plant-Microbe Interact 13:942–950
 36. Janni M, Sella L, Favaron F, Blechl A, De Lorenzo G, D'Ovidio R (2008) The expression of a bean PGIP in transgenic wheat confers increased resistance to the fungal pathogen *Bipolaris sorokiniana*. Mol Plant-Microbe Interact 21:171–178
 37. Desiderio A, Aracri B, Leckie F, Mattei B, Salvi G, Tigelaar H, Van Roekel JS, Baulcombe DC, Melchers LS, De Lorenzo G, Cervone F (1997) Polygalacturonase-inhibiting proteins (PGIPs) with different specificities are expressed in *Phaseolus vulgaris*. Mol Plant-Microbe Interact 10:852–860
 38. Hu X, Bidney DL, Yalpani N, Duvick JP, Crasta O, Folkerts O, Lu G (2003) Overexpression of a gene encoding hydrogen peroxide-generating oxalate oxidase evokes defense responses in sunflower. Plant Physiol 133:170–181
 39. Walz A, Zingen-Sell I, Loeffler M, Sauer M (2008) Expression of an oxalate oxidase gene in tomato and severity of disease caused by *Botrytis cinerea* and *Sclerotinia sclerotiorum*. Plant Pathol 57:453–458
 40. Liang H, Zheng J, Duan XY, Sheng BQ, Jia SG, Wang DW, Ouyang JW, Li JY, Li LC, Tian WZ, Hain R, Jia X (2000) A transgenic wheat with a stilbene synthase gene resistant to powdery mildew obtained by biolistic method. Chin Sci Bull 45:634–638
 41. Schneider M, Droz E, Malnoe P, Chatot C, Bonnel E, Metrauz J (2002) Transgenic potato plants expressing oxalate oxidase have increased resistance to oomycete and bacterial pathogens. Potato Res 45:177–185
 42. Sanford JC, Johnston SA (1985) The concept of parasite-derived resistance-deriving resistance genes from the parasite's own genome. J Theor Biol 113:395–405
 43. Wang GL, Song WY, Ruan DL, Sideris S, Ronald PC (1996) The cloned gene, *Xa21*, confers resistance to multiple *Xanthomonas oryzae* pv. *oryzae* isolates in transgenic plants. Mol Plant-Microbe Interact 9:850–855
 44. Datta SK (2004) Rice biotechnology: a need for developing countries. AgBioForum 7:31–35
 45. Zhang J, Li X, Jiang G, Xu Y, He Y (2006) Pyramiding of *Xa7* and *Xa21* for the improvement of disease resistance to bacterial blight in hybrid rice. Plant Breed 125:600–605
 46. Yoshimura S, Yamanouchi U, Katayose Y, Toki S, Wang ZX, Kono I, Kurata N, Yano M, Iwata N, Sasaki T (1998) Expression of *Xa1*, a bacterial blight-resistance gene in rice, is induced by bacterial inoculation. Proc Natl Acad Sci USA 95:1663–1668
 47. Zhao BY, Lin XH, Poland J, Trick H, Leach J, Hulbert S (2005) From the covers: A maize resistance gene functions against bacterial streak disease in rice. Proc Natl Acad Sci USA 102:15383–15388
 48. Tai TH, Dahlbeck D, Clark ET, Gajiwala P, Pasion R, Whalen MC, Stall RE, Staskawicz BJ (1999) Expression of the *Bs2* pepper gene confers resistance to bacterial spot disease in tomato. Proc Natl Acad Sci USA 23:14153–14158
 49. Mysore KS, D'Ascenzo MD, He X, Martin G (2003) Overexpression of the disease resistance gene *Pto* in tomato induces gene expression changes similar to immune responses in human and fruitfly. Plant Physiol 132:1901–1912
 50. Arce P, Moreno M, Gutierrez M, Gebauer M, Dell'Orto P, Torres H, Acuna I, Oliger P, Venegas A, Jordana X, Kalazich J, Hluigie L (1999) Enhanced resistance to bacteria infection by *Erwinia carotovora* subsp. *atroseptica* in transgenic potato plants expression the *attacin* or the *cecropin* SB-37 genes. Am J Potato Res 76:169–177
 51. Norelli JL, Mills JZ, Jensen LA, Momol MT, Ko K, Aldwinckle HS (1998) Genetic engineering of apple for increased resistance to fire blight. Acta Hort 484:541–546
 52. Mentag R, Luckevich M, Morency MJ, Seguin A (2003) Bacterial disease resistance of transgenic hybrid poplar expressing the synthetic antimicrobial peptide D4E1. Tree Physiol 23:405–411
 53. Montesinos E (2007) Antimicrobial peptides and plant disease control. FEMS Microbiol Lett 270:1–11
 54. Reynold JP, Mourgues F, Norelli J, Aldwinckle HS, Brisset MN, Chevreau E (1999) First evidence for improved resistance to fire blight in transgenic pear expressing the *attacin* E gene from *Hyalophora cecropia*. Plant Sci 149:23–31
 55. Malnoy M, Venisse JS, Brisset MN, Chevreau E (2003) Expression of bovine lactoferrin cDNA confers resistance to *Erwinia amylovora* in transgenic pear. Mol Breed 12:231–244
 56. Norelli JL, Mills JZ, Momol MT, Aldwinckle HS (1999) effect of *cecropin*-like transgenes on fire blight resistance of apple. Acta Hort 489:273–278

57. Ko K, Norelli JL, Reynold JP, Boreszja-Wysocka E, Brown SK, Aldwinckle HS (2000) Effect of untranslated leader sequence of AMV RNA 4 and signal peptide of pathogenesis-related protein 1b on attacin gene expression, and resistance to fire blight in transgenic apple. *Biotech Lett* 22:373–381
58. Düring K, Porsch P, Fladung M, Lörz H (1993) Transgenic potato plants resistant to the phytopathogenic bacterium *Erwinia carotovora*. *Plant J* 3:587–598
59. Ko K (1999) Attacin and T4 lysozyme transgenic 'Galaxy' apple. Regulation of transgene expression and plant resistance to fire blight (*Erwinia amylovora*). PhD dissertation, Cornell University, New York
60. Mirkov TE, Fitzmaurice LC (1995) Protection of plants against plant pathogens. US Patent No 5422108
61. Allefs JJHM, de Jong ER, Florack DEA, Hoogendoorn J, Stiekema WJ (1996) *Erwinia* soft rot resistance of potato cultivars expressing antimicrobial peptide tachyplesin. *Mol Breed* 2: 97–105
62. Wu G, Shortl BJ, Lawrence EB, Levine EB, Fitzsimmons KC, Shah DM (1995) Disease resistance conferred by expression of a gene encoding H₂O₂-generating glucose oxidase in transgenic potato plants. *Plant Cell* 7:1357–1368
63. Dong YH, Wang L, Xu JL, Zhang HB, Zhang XF, Zhang LH (2001) Quenching quorum-sensing-dependent bacterial infection by an *N*-acyl homoserine lactonase. *Nature* 411:813–817
64. Herrera-Estrella L, Simpson J (1995) Genetically engineered resistance to bacterial and fungal pathogens. *World J Microbiol Biotechnol* 11:383–392
65. Krastanova SV, Sekiya M, Holden MR, Xue B, Pang S-Z, Momol EA, Gonsalves D, Burr TJ (2010) Transformation of grape rootstocks with a truncated *Agrobacterium* virE2 gene confers resistance to crown gall disease. *Trans Res* 19:949–958
66. Citovsky V, Warnick D, Zambryski P (1994) Nuclear import of *Agrobacterium* VirD2 and VirE2 proteins in maize and tobacco. *Proc Natl Acad Sci USA* 91:3210–3214
67. Escobar MA, Civerolo EL, Summerfelt KR, Dandekar AM (2001) RNAi-mediated oncogene silencing confers resistance to crown gall tumorigenesis. *Proc Natl Acad Sci USA* 98:13437–13442
68. Lin WC, Lu CF, Wu JW, Cheng ML, Lin YM, Yang NS, Black L, Green SK, Wang JF, Cheng CP (2004) Transgenic tomato plants expressing the Arabidopsis NPR1 gene display enhanced resistance to a spectrum of fungal and bacterial diseases. *Trans Res* 13:567–581
69. Makandar R, Essig JS, Schapaugh MA, Trick HN, Shah J (2006) Genetically engineered resistance to Fusarium head blight in wheat by expression of Arabidopsis NPR1. *Mol Plant Microbe Interact* 19:123–129
70. Meur G, Budatha M, Srinivasan T, Kumar KRR, Gupta AD, Kirti PB (2008) Constitutive expression of Arabidopsis NPR1 confers enhanced resistance to the early instars of *Spodoptera litura* in transgenic tobacco. *Physiol Plant* 133:765–775
71. Wally O, Jayaraj J, Punja ZK (2009) Broad-spectrum disease resistance to necrotrophic and biotrophic pathogens in transgenic carrots (*Daucus carota* L.) expressing an Arabidopsis NPR1 gene. *Planta* 231:131–141
72. Chen MS, Fitzgerald HA, Yadav RC, Canlas PE, Dong XN, Ronald PC (2001) Evidence for a disease-resistance pathway in rice similar to the NPR1-mediated signaling pathway in Arabidopsis. *Plant J* 27:101–113
73. Chen M, Fitzgerald HA, Canlas PE, Navarre DA, Ronald PC (2005) Overexpression of a rice NPR1 homolog leads to constitutive activation of defense response and hypersensitivity to light. *Mol Plant-Microbe Interact* 18:511–520
74. Malnoy M, Jin Q, Borszja-Wysocka EE, He SY, Aldwinckle HS (2007) Overexpression of the apple MpNPR1 gene confers increased disease resistance in *Malus x domestica*. *Mol Plant-Microbe Interact* 20:1568–1580
75. Navarro L, Dunoyer P, Jay F, Arnold B, Dharmasiri N, Estelle M, Voinnet O, Jones JD (2006) A plant miRNA contributes to antibacterial resistance by repressing auxin signaling. *Science* 312:436–439
76. Agorio A, Vera P (2007) ARGONAUTE4 is required for resistance to *Pseudomonas syringae* in Arabidopsis. *Plant Cell* 19:3778–3790
77. Baulcombe D (2007) Amplified silencing. *Science* 315:199–200
78. Eamens A, Wang M-B, Smith NA, Waterhouse PM (2008) RNA silencing in plants: yesterday, today and tomorrow. *Plant Physiol* 147:456–468
79. Lin S-S, Henriques R, Wu H-W, Niu Q-W, Yeh S-D, Chua N-H (2007) Strategies and mechanisms of plant virus resistance. *Plant Biotech Rep* 1:125–134
80. Obbard DJ, Gordon KH, Buck AH, Jiggins FM (2009) The evolution of RNAi as a defence against viruses and transposable elements. *Phil Trans R Soc B* 364:99–115
81. Prins M, Laimer M, Noris E, Schubert J, Wassenegger M, Tepfer M (2008) Strategies for antiviral resistance in transgenic plants. *Mol Plant Pathol* 9:73–83
82. Voinnet O (2008) Post-transcriptional RNA silencing in plant-microbe interactions: a touch of robustness and versatility. *Curr Opin Plant Biol* 11:464–470
83. Powell Abel P, Nelson RS, De B, Hoffman N, Rogers SG, Fraley RT, Beachy RN (1986) Delay of disease development in transgenic plants that express the tobacco mosaic virus coat protein gene. *Science* 232:738–743
84. Nelson RS, McCormick SM, Delannay X, Dube P, Layton J, Anderson EJ, Kaniewska M, Proksch RK, Horsch RB, Rogers SG, Fraley RT, Beachy RN (1988) Virus tolerance, plant growth and field performance of transgenic tomato plants expressing coat protein from tobacco mosaic virus. *Bio/Technology* 6:403–409
85. Gonsalves D, Chee P, Provvidenti R, Seem R, Slightom JL (1992) Comparison of coat protein-mediated and genetically-derived resistance in cucumbers to infection by *Cucumber mosaic virus* under field conditions with natural challenge inoculations by vectors. *Bio/Technology* 10:1562–1570
86. Lawson C, Kaniewski W, Haley L, Rozman R, Newell C, Sanders P, Tumer N (1990) Engineering resistance to mixed virus infection in a commercial potato cultivar: resistance of *Potato virus X* and *Potato virus Y* in transgenic Russet Burbank. *Bio/Technology* 8:127–134

87. Kaniewski W, Lawson C, Sammons B, Haley L, Hart J, Delannay X, Tumer N (1990) Field resistance of transgenic Russet Burbank potato to effects of infection by *Potato virus X* and *Potato virus Y*. *Bio/Technology* 8:750–754
88. Tricoli DM, Carney KJ, Russell PF, McMaster JR, Groff DW, Hadden KC, Himmel PT, Hubbard JP, Boeshore ML, Quemada HD (1995) Field evaluation of transgenic squash containing single or multiple virus coat protein gene constructs for resistance to *Cucumber mosaic virus*, *Watermelon mosaic virus 2*, and *Zucchini yellow mosaic virus*. *Bio/Technology* 13:1458–1465
89. Fuchs M, Gonsalves D (1995) Resistance of transgenic squash Pavo ZW-20 expressing the coat protein genes of zucchini yellow mosaic virus and watermelon mosaic virus 2 to mixed infections by both potyviruses. *Bio/Technology* 13:1466–1473
90. Fuchs M, Tricoli DM, McMaster JR, Carney KJ, Schesser M, McFerson JR, Gonsalves D (1998) Comparative virus resistance and fruit yield of transgenic squash with single and multiple coat protein genes. *Plant Dis* 82:1350–1356
91. Lindbo JA, Dougherty WG (2005) Plant Pathology and RNAi: a brief history. *Annu Rev Phytopathol* 43:191–204
92. Lindbo JA, Silva-Rosales L, Proebsting WM, Dougherty WG (1993) Induction of a highly specific antiviral state in transgenic plants: implications for regulation of gene expression and virus resistance. *Plant Cell* 5:1749–1759
93. Waterhouse PM, Graham MW, Wang MB (1998) Virus resistance and gene silencing in plants can be induced by simultaneous expression of sense and antisense RNA. *Proc Natl Acad Sci USA* 95:13959–13964
94. Hamilton AJ, Baulcombe DC (1999) A species of small antisense RNA in post-transcriptional gene silencing in plants. *Science* 286:950–952
95. Hamilton AJ, Voinnet O, Chappell L, Baulcombe DC (2002) Two classes of short interfering RNA in RNA silencing. *EMBO J* 21:4671–4679
96. Deleris A, Gallego-Bartolome J, Bao J, Kasschau KD, Carrington JC, Voinnet O (2006) Hierarchical action and inhibition of plant dicer-like proteins in antiviral defense. *Science* 313:68–71
97. Hannon GJ (2002) RNA interference. *Nature* 418:244–251
98. Smith NA, Singh SP, Wang M-B, Stoutjesdijk P, Green A, Waterhouse PM (2000) Total silencing by intron-spliced hairpin RNAs. *Nature* 407:319–320
99. Wesley V, Helliwell CA, Smith NA, Wang MB, Rouse DT, Liu Q, Gooding PS, Singh SP, Abbott D, Stoutjesdijk PA, Robinson SP, Gleave AP, Green AG, Waterhouse PM (2001) Construct design for efficient, effective and high-throughput gene silencing in plants. *Plant J* 27:581–590
100. Bucher E, Lohuis D, van Poppel PMJA, Geerts-Dimitriadou C, Goldbach R, Prins M (2006) Multiple virus resistance at a frequency using a single transgene construct. *J Gen Virol* 87:3697–3701
101. Niu Q-W, Lin S-S, Reyes JL, Chen K-C, Wu H-W, Yeh S-D, Chua N-H (2006) Expression of artificial microRNAs in transgenic *Arabidopsis thaliana* confers virus resistance. *Nat Biotechnol* 11:1420–1428
102. Qu F, Ye X, Morris TJ (2008) Arabidopsis DRB4, AGO1, AGO7, and RDR6 participate in a DCL4-initiated antiviral RNA silencing pathway negatively regulated by DCL1. *Proc Natl Acad Sci USA* 105:14732–14737
103. Rudolph C, Schreier PH, Uhrig JF (2003) Peptide-mediated broad-spectrum plant resistance to tospoviruses. *Proc Natl Acad Sci USA* 100:4429–4434
104. Boonrod K, Galetzka D, Nagy PD, Conrad U, Krczal G (2004) Single-chain antibodies against a plant viral RNA-dependent RNA polymerase confer virus resistance. *Nat Biotechnol* 22:856–862
105. Nölke G, Fischer R, Shillberg S (2004) Antibody-based pathogen resistance in plants. *J Plant Pathol* 86:5–17
106. Nölke R, Cobanov P, Uhde-Holzem K, Reustle G, Fischer R, Shillberg S (2009) *Grapevine fanleaf virus* (GFLV)-specific antibodies confer GFLV and *Arabidopsis mosaic virus* (ArMV) resistance in *Nicotiana benthamiana*. *Mol Plant Pathol* 10:41–49
107. Schillberg S, Zimmermann S, Findlay K, Fischer R (2000) Plasma membrane display of anti-viral single chain Fv fragments confers resistance to *Tobacco mosaic virus*. *Mol Breed* 6:317–326
108. Tavladoraki P, Benvenuto E, Trinca S, De Martinis D, Cattaneo A, Galeffi P (1993) Transgenic plants expressing a functional single-chain Fv antibody are specifically protected from virus attack. *Nature* 366:469–472
109. Lodge JK, Kaniweski WK, Tumer NE (1993) Broad-spectrum virus resistance in transgenic plants expressing pokeweed antiviral protein. *Proc Natl Acad Sci USA* 90:7089–7093
110. Gutierrez-Campos R, Torres-Acosta JA, Saucedo-Arias LJ, Gomez-Lim MA (1999) The use of cysteine proteinase inhibitors to engineer resistance against potyviruses in transgenic tobacco plants. *Nat Biotechnol* 17:1223–1226
111. Mitra A, Higgins DW, Langenberg WG, Nie H, Sengupta DN, Silverman RH (1996) A mammalian 2–5A system functions as an antiviral pathway in transgenic plants. *Proc Natl Acad Sci USA* 93:6780–6785
112. Ogawa T, Hori T, Ishida I (1996) Virus-induced cell death in plants expressing the mammalian 2', 5', oligoadenylate system. *Nat Biotechnol* 14:1566–1569
113. Gonsalves D (1998) Control of *Papaya ringspot virus* in papaya: a case study. *Annu Rev Phytopathol* 36:415–437
114. Johnson SR, Strom S, Grillo K (2007) Quantification of the impacts on US agriculture of biotechnology-derived crops planted in 2006. National Center for Food and Agriculture policy. <http://www.ncfap.org>
115. James C (2010) Global status of commercialized biotech/GM crops. The first thirteen years, 1996 to 2008. <http://www.isaaa.org/resources/publications/briefs/39/executivesummary/default.html>
116. Stone R (2008) China plans \$3.5 billion GM crops initiative. *Science* 321:1279
117. Kaniewski WK, Thomas PE (2004) The potato story. *Agric Biol Forum* 7:41–46

118. Capote N, Pérez-Panadés J, Monzó C, Carbonell E, Urbaneja A, Scorza R, Ravelonandro M, Cambra M (2007) Assessment of the diversity and dynamics of *Plum pox virus* and aphid populations in transgenic European plums under Mediterranean conditions. *Trans Res* 17:367–377
119. Hily JM, Scorza R, Malinowski T, Zawadzka B, Ravelonandro M (2004) Stability of gene silencing-based resistance to *Plum pox virus* in transgenic plum (*Prunus domestica* L.) under field conditions. *Transgenic Res* 5:427–436
120. Malinowski T, Cambra M, Capote N, Zawadzka B, Gorris MT, Ravelonandro M (2006) Field trials of plum clones transformed with the *Plum pox virus* coat protein (PPV-CP) gene. *Plant Dis* 90:1012–1018
121. Zagrai I, Capote N, Ravelonandro M, Ravelonandro M, Cambra M, Zagrai L, Scorza R (2008) *Plum pox virus* silencing of C5 transgenic plums is stable under challenge inoculation with heterologous viruses. *J Plant Pathol* 90:S1.63–S1.71
122. Bech R (2007) Finding of no significant impact and decision notice. Determination of nonregulated status for C5 plum resistant to plum pox virus. Animal and Plant Health Inspection Service. http://www.aphis.usda.gov/brs/aphisdocs/04_26401p_com.pdf
123. Davis MJ (2004) Petition for Determination of Nonregulated status for the X17-2 line of papaya: A *Papaya ringspot virus*-resistant papaya. Animal and Plant Health Inspection Service. http://www.aphis.usda.gov/brs/aphisdocs/04_33701p.pdf
124. Shea K (2009) University of Florida; determination of nonregulated status for papaya genetically engineered for resistance to the *Papaya ringspot virus*. Animal and Plant Health Inspection Service. Fed Regist 74:45163–45164. <http://edocket.access.gpo.gov/2009/pdf/E9-21092.pdf>
125. Anand A, Zhou T, Trick HN, Gill BS, Bockus WW, Muthukrishnan S (2003) Greenhouse and field testing of transgenic wheat plants stably expressing genes for thaumatin-like protein, chitinase and glucanase against *Fusarium graminearum*. *J Exp Bot* 54:1101–1111
126. Chenault K, Melouk H, Payton M (2005) Field reaction to *Sclerotinia* blight among transgenic peanut lines containing antifungal genes. *Crop Sci* 45:511–515
127. Punja Z, Raharjo S (1996) Response of transgenic cucumber and carrot plants expressing different chitinase enzymes to inoculation with fungal pathogens. *Plant Dis* 80:999–1005
128. Zhang G, Chen M, Li L, Xu Z, Chen X, Jiaming Guo J, Youzhi M (2009) Overexpression of the soybean *GmERF3* gene, an AP2/ERF type transcription factor for increased tolerances to salt, drought, and diseases in transgenic tobacco. *J Exp Bot* 60:3781–3796
129. Joosten MH, Cozijnsen TJ, De Wit PJ (1994) Host resistance to fungal tomato pathogen lost by a single base pair change in an avirulence gene. *Nature* 367:384–386
130. Gassmann W, Dahlbeck D, Chesnokova O, Minsavage G, Jones J, Stakawicz B (2000) Molecular evolution of virulence in natural field strains of *Xanthomonas campestris* pv. *vesicatoria*. *J Bacteriol* 182:7053–7059
131. Ellis JG, Lawrence GJ, Luck JE, Dodds PN (1999) Identification of regions in alleles of the flax rust resistance gene *L* that determine differences in gene-for-gene specificity. *Plant Cell* 11:495–506
132. Jia Y, McAdams S, Bryan E, Hershey H, Valent B (2000) Direct interaction of the resistance gene and avirulence gene products confers rice blast resistance. *EMBO J* 14:4004–4014
133. Gurr SJ, Rushton PJ (2005) Engineering plants with increased disease resistance: how are we going to express it? *Trends Biotechnol* 23:283–290
134. Eulgen T (2005) Regulation of the *Arabidopsis* defence transcriptome. *Trends Plant Sci* 10:71–78
135. Rushton P, Reinstädler A, Lipka V, Lippok B, Somssich I (2002) Synthetic plant promoters containing defined regulatory elements provide novel insights into pathogen- and wound-induced signaling. *Plant Cell* 14:749–762
136. Kovalchuk N, Li M, Wittek F, Reid N, Singh R, Shirley N, Ismagul A, Eliby S, Johnson A, Milligan A, Hrmova M, Langridge P, Lopato S (2010) Defensin promoters as potential tools for engineering disease resistance in cereal grains. *Plant Biotechnol J* 8:47–64
137. Chan Y, Prasad V, Sanjaya K, Chen K, Liu P, Chan M, Cheng C (2005) Transgenic tomato plants expressing an *Arabidopsis* thionin (*Thi2.1*) driven by fruit-inactive promoter battle against phytopathogenic attack. *Planta* 221:386–393
138. Calo L, Garcia I, Gotor C, Romero LC (2006) Leaf hairs influence phytopathogenic fungus infection and confer an increased resistance when expressing a *Trichoderma* a-1, 3-glucanase. *J Exp Bot* 57:3911–3920

Books and Reviews

- Charkowski AO (2007) Potential disease control strategies revealed by genome sequencing and functional genetics of plant pathogenic bacteria. In: Punja ZK, De Boer SH, Sanfaçon H (eds) *Biotechnology and plant disease management*. CAB International, Cambridge, MA, pp 402–497
- Csroba T, Pantaleo V, Burgyán J (2009) RNA silencing: an antiviral mechanism. In: Loebeinstein G, Carr JP (eds) *Advances in virus research*, vol 75, natural and engineered resistance to plant viruses. Academic, San Diego, CA, pp 35–72
- Diaz-Pendón JA, Ding S-W (2008) Direct and indirect roles of viral suppressors of RNA silencing in pathogenesis. *Annu Rev Phytopathol* 46:303–326
- Knox RE, Clarke FR (2007) Molecular breeding approaches for enhanced resistance against fungal pathogens. In: Punja ZK, De Boer SH, Sanfaçon H (eds) *Biotechnology and plant disease management*. CAB International, Cambridge, MA, pp 321–357
- Misra S, Bhargava A (2007) Application of cationic antimicrobial peptides for management of plant diseases. In: Punja ZK, De Boer SH, Sanfaçon H (eds) *Biotechnology and plant disease management*. CAB International, Cambridge, MA, pp 301–320

- Rommens CM, Kishore GM (2000) Exploiting the full potential of disease-resistance genes for agricultural use. *Curr Opin Biotechnol* 11:120–125
- Ruiz-Ferrer V, Voinnet O (2009) Roles of plant small RNAs in biotic stress responses. *Annu Rev Plant Biol* 60:485–510
- Stuiver MH (2006) Engineering fungal resistance in crops. In: Halford N (ed) *Plant biotechnology: current and future applications of genetically modified crops*. Wiley, West Sussex, England, pp 225–239
- Stuiver MH, Custers JH (2000) Engineering disease resistance in plants. *Nature* 411:865–868
- Suzuki JY, Tripathi S, Gonsalves D (2007) Virus-resistant transgenic papaya: commercial development and regulatory and environmental issues. In: Punja ZK, De Boer SH, Sanfaçon H (eds) *Biotechnology and plant disease management*. CAB International, Cambridge, MA, pp 436–461
- Xing T (2007) Signal transduction pathways and disease resistant genes and their applications to fungal disease control. In: Punja ZK, De Boer SH, Sanfaçon H (eds) *Biotechnology and plant disease management*. CAB International, Cambridge, MA, pp 1–15
- Yang CH, Yang S (2007) Modulating quorum sensing and type III secretion systems in bacterial plant pathogens for disease management. In: Punja ZK, De Boer SH, Sanfaçon H (eds) *Biotechnology and plant disease management*. CAB International, Cambridge, MA, pp 16–57

Transgenic Crops, Environmental Impact

ANGHARAD M. R. GATEHOUSE¹, MARTIN G. EDWARDS¹,
NATALIE FERRY^{1,2}, MICHEAL D. K. OWEN³

¹School of Biology, Newcastle Institute for Research on Sustainability, Newcastle University, Newcastle upon Tyne, UK

²School of Environment and Life Sciences, University of Salford, Salford, UK

³Department of Agronomy, Iowa State University, Ames, IA, USA

Article Outline

Glossary

Definition of the Subject

Introduction

Insect-Resistant Transgenic Crops

Herbicide-Tolerant/Resistant Transgenic Crops

Environmental Impact of Transgenic Crops

Future Directions

Conclusion

Bibliography

Glossary

***Bt*-expressing transgenic crops**

Insect resistance crops

Novel mechanisms for insect resistance

Herbicide-tolerant transgenic crops

Food security

Sustainability

Risk assessment

Environmental impact

Impact on beneficial insects

Gene flow

Definition of the Subject

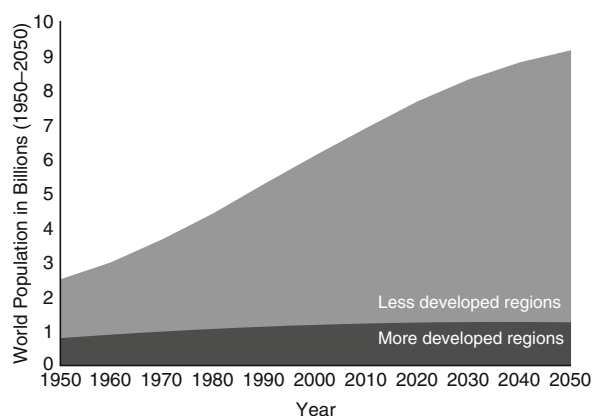
Agriculture is the production of food and commodities through farming. Since its dawn some 10,000 years ago, humankind has had an ever increasing impact on the environment. With the increasing intensification of agriculture, particularly during the twentieth century, this impact has become even more pronounced, often with undesirable or unacceptable consequences, including water pollution, soil erosion, and loss of habitat, often accompanied by a loss of biodiversity. Pests, particularly weedy plants, demonstrate a remarkable ability to adapt to agricultural production systems. The practice of growing monocultures, typically used in intensive agriculture, increases the number of pests; these are currently predominantly controlled through use of pesticides. However, with increasing exposure to pesticides, many pest populations are evolving resistance to these compounds. An additional problem encountered with many pesticides, and particularly insecticides, is their non-target effects on beneficial insects. Transgenic crops expressing genes conferring either resistance to insect pests and/or herbicides are becoming increasingly more widely grown. However, in many parts of the world, before such crops are commercialized, they have to be assessed for their potential environmental impact which measures the impact they will have on non-target species and their ability to outcompete native species. Environmental risk assessments cover both the transgenic crop

concerned and the potential impacted environment. The assessment process includes evaluation of the characteristics of the crop and its effect and stability in the environment, combined with ecological characteristics of the environment in which the introduction will take place. The assessment also includes unintended consequences that could result from the insertion of the new gene. For all transgenic crops, the direct and indirect effects they have on non-target organisms must be considered. Furthermore, issues relating to potential gene flow to other non-transgenic cultivars and/or wild near-relatives also must be addressed.

Introduction

Food security figures are high on both the political and social agenda [1]. This is not surprising given that the global population increased fourfold during the last century, with current estimates suggesting that it will reach approximately 9.2 billion by 2050 (Fig. 1). However, this increase has primarily been seen in the developing regions of the world, with population remaining relatively static in the developed world. As expected, it is those countries least able that need to significantly increase food production, with an estimated increase in productivity of 300% for Africa and 80% for Latin America; even so, it is estimated that North America will need to increase productivity by some 30% [2]. Without an increase in “farm productivity,” this required increase in food production equates to an additional 1.6 billion hectares of arable land by 2050, a scenario that is becoming increasingly less sustainable or, indeed, achievable. An immediate priority for agriculture is thus to achieve increased crop yields in a sustainable and cost-effective way in order to avoid a Malthusian disaster.

The concept of utilizing a transgenic (biotech) approach was realized in the mid-1990s with the commercial introduction of genetically modified crops. In 2009, it was estimated that some 14 million farmers planted 134 million hectares (330 million acres) of biotech crops in 25 countries, representing a 7% increase over the preceding year [3]. Interestingly, 90% of these farmers were small and resource-poor farmers from developing countries. In economic terms, the global market value for biotech crops in 2009 was US\$10.5 billion; this figure represents 20%



Transgenic Crops, Environmental Impact. Figure 1

Project world population increase from 1950 to 2050 showing an overall fourfold increase in population. During this period, population growth in the more developed regions of the world are predicted to remain moderately static

of the US\$52.2 billion global crop protection market and 30% of US\$34 billion commercial seed market. Of this US\$10.5 billion biotech crop market, approximately 50% was accounted for by biotech maize, with soybean, cotton, and canola (oilseed rape) accounting for 37.2%, 10.5%, and 3%, respectively.

Although transgenic crops have only been available since the 1990s, they have dramatically changed the face of world agriculture and are likely to represent the most rapidly adopted technology in agriculture [4]. This trend is largely attributed to one herbicide, glyphosate (*N*-(phosphonomethyl)glycine), and crops that are genetically modified to have selectivity (resistance) to the glyphosate when the herbicide is applied topically to them. However, many new transgenic crops have been introduced that include multiple transgenic traits, thus providing greater value to agriculture. The reported increased economic profitability attributable to transgenic crops is a major factor that provides the impetus for the increasing adoption of these crops worldwide, both in developing as well as industrialized nations [5]. Despite their benefits, not least in increasing crop yield in a more sustainable and environmentally benign way compared to intensive agriculture (Tables 1 [6] and 2 [6–20]), neither the technology, nor the crops, are universally perceived as of benefit to humankind. There are fears that all

Transgenic Crops, Environmental Impact. Table 1 Average farm level agronomic and economic effects of *Bt* crops

Country	Insecticide reduction (%)	Increase in effective yield (%)	Increase in gross margin (US\$/ha)
	<i>Bt cotton</i>	<i>Bt cotton</i>	<i>Bt cotton</i>
Argentina	47	33	23
Australia	48	0	66
China	65	24	470
India	41	37	135
Mexico	77	9	295
South Africa	33	22	91
USA	36	10	58
	<i>Bt maize</i>	<i>Bt maize</i>	<i>Bt maize</i>
Argentina	0	9	20
Philippines	5	34	53
South Africa	10	11	42
Spain	63	6	70
USA	8	5	12

Source: Zilberman et al. (2010) [6] (Courtesy of *Choices* and the Agricultural & Applied Economics Association)

questions of risk attributable to the wide-scale growing of such crops have not been adequately addressed, or perhaps even identified [21]. While many of these fears may be unfounded, public debate concerning potential environmental and public health risks is crucial.

This chapter will provide a brief overview of first-generation transgenic crops. Specifically, it will discuss the environmental impact of the wide-scale growing of such crops both in terms of their potential non-target effects and the potential for gene flow.

Insect-Resistant Transgenic Crops

Insect-resistant transgenic crops, also known as Biotech crops, were first commercialized in the USA during the

Transgenic Crops, Environmental Impact. Table 2 Summary of primary studies on the effects of herbicide-tolerant (HT) crops on yields

Crop/Reference	Data source	Effect on yields
<i>Herbicide-tolerant soybeans</i>		
[7]	Experiments	Same
[8]	Experiments	Increase
[9]	Experiments	Increase
[10]	Survey	Increase
[11]	Survey	Small increase
[12]	Survey	Small decrease
[13]	Survey	Same
[14]	Survey	Increase
[15]	Survey	Same
<i>Herbicide-tolerant cotton</i>		
[16]	Experiments	Same
[17]	Experiments	Same
[18]	Experiments	Same
[19]	Experiments	Same
[20]	Survey	Increase

Source: Zilberman et al. (2010) [6] (Courtesy of *Choices* and the Agricultural & Applied Economics Association)

mid-1990s with the introduction of genetically modified corn (maize), potato, and cotton plants expressing genes encoding the entomocidal δ -endotoxin from *Bacillus thuringiensis* (*Bt*; also known as crystal/Cry proteins). In terms of *Bt*-expressing maize and cotton, the current market is estimated at US\$2.3 billion and US\$0.9 billion, respectively. Detailed knowledge on the mode of action of these Cry proteins is not only essential to optimize their efficacy against target insects, but also to predict the potential for non-target effects on beneficial insects. On ingestion, *Bt* toxins are solubilized in the midgut where they are proteolytically cleaved at the N-terminal to a 65–70-kDa truncated (active) form. The active molecules then exert their pathological effects by binding to a specific receptor (s) in the midgut epithelial cells and insert into the membrane where they form pores, this results in cell

death by colloid osmotic lysis, followed by death of the insect [22]. In most commercial crop varieties, these Cry proteins are expressed in the active form and as such differ from those used in biopesticide formulations where the Cry proteins are present as protoxins.

Early commercial varieties of insect-resistant transgenic crops expressed single crystal (Cry) proteins with specific activity against lepidopteran pests as illustrated by Bollgard® cotton expressing Cry1Ac developed by Monsanto and Attribute® maize expressing Cry1Ab developed by Syngenta. Subsequently, other lepidopteran-active *Bt* toxins, such as Cry1F and Cry2Ab2, were introduced, and often presented as pyramided genes in a single variety (Widestrike® cotton expressing both Cry1F + Cry1Ac developed by Dow Agrosciences and Bollgard II® cotton expressing Cry1Ac + Cry2Ab2 developed by Monsanto). Transgenic crops, particularly maize, expressing Cry3 proteins to protect against coleopteran pests such as chrysomelid root-worms have also been commercialized (e.g., Monsanto's Yieldgard Rootworm® maize expressing Cry3Bb1, Dow Agrosciences' Herculex RW® maize expressing Cry34Ab1 and Cry35Ab1, stacked with a HT gene, and Syngenta's Agrisure RW® maize expressing a modified version of Cry3A). In 2009, SmartStax, a novel biotech maize containing eight different genes for insect and herbicide resistance, was granted approval in the USA. Furthermore, Syngenta has recently launched Agrisure Viptera trait-stacked corn, the first commercially available variety to exploit a non-cry *Bt* protein (Vip3) for the provision of multiple pest resistance. In China, *Bt* cotton cultivars expressing Cry1Ac together with a modified cowpea trypsin inhibitor (CpTI) were commercially released in 2000 [23], and in 2005, accounted for approximately 15% of the cotton crop [24]. The major reason for co-expressing *Bt* and CpTI in cotton was to reduce the likelihood of the insects becoming resistant to this cultivar, thus extending its effective life.

While it is clear that *Bt*-expressing crops have made a significant beneficial impact to global agriculture, not least in terms of pest reduction, improved quality, and increased yield (Table 1 [6]), it is also becoming increasingly evident that there is a need to develop alternative strategies, not least because of the potential for pest populations to evolve resistance [25], and to the lack of effective control of homopteran pests.

Recently, crops expressing vegetative insecticidal proteins (VIPs) have been commercialized (see above). However, alternative approaches based on the use of plant-derived or animal-derived genes, including those from insects (such as those encoding immuno-suppressive proteins), are being investigated and developed. More recently, the potential to identify and exploit endogenous resistance genes using functional genomics and the use of RNAi are actively being investigated [26–30]. However, if these novel approaches are to play a useful role in crop protection, it is desirable that they do not have a negative impact on beneficial organisms at higher trophic levels since this would inevitably affect agro-ecosystem function. This is equally true for *Bt*-expressing crops.

Herbicide-Tolerant/Resistant Transgenic Crops

From the first commercialization of transgenic crops, to date, herbicide tolerance has consistently been the dominant trait and, as a consequence of its rapid and widespread adoption, has made an immeasurable change to global agriculture. In 2009, herbicide tolerance deployed in soybean, maize (corn), canola (oilseed rape), cotton, sugarbeet, and alfalfa occupied 62% (83.6 million hectares) of the global biotech area of 134 million hectares. Furthermore, in 2009, the stacked double and triple traits occupied an area of 28.7 million hectares (equivalent to 21% of global biotech crop area), much larger than the insect-resistant varieties which occupied 21.7 million hectares (15%). Interestingly, stacked trait products and herbicide-tolerant products grew at the same rate of 6%, while insect resistance as a trait grew at 14% [31].

While there are a number of herbicide-tolerant transgenic crops that have been developed to several herbicides with different modes of phytotoxic action, the primary influence in world agriculture is glyphosate [32, 33]. This preference for glyphosate is based on (1) the target site, (2) the ability of this compound to translocate in plants, and (3) the inability of plants to rapidly detoxify it. Glyphosate controls weeds by inhibiting 5-enolpyruvylshikimate-3-phosphate synthase (EPSPS; EC 2.5.1.19), a key enzyme in the shikimate biosynthetic pathway which is necessary for the production of the aromatic amino acids, auxin, phytoalexins, folic acid, lignin, plastoquinones, and many other secondary plant

products. Over 30% of the carbon fixed by plants may pass through this pathway. Inhibition of EPSPS by glyphosate deregulates the pathway, leading to even more carbon flowing through the pathway with accumulation of shikimate and shikimate-3-phosphate. Up to 16% of the plant's dry matter can accumulate as shikimate. Glyphosate occupies the binding site on EPSPS for phosphoenolpyruvate, a substrate of EPSPS, by mimicking an intermediate state of the enzyme-substrate complex. There are two forms of EPSPS in nature, EPSPS I, which is found in plants, fungi, and most bacteria, and is sensitive to glyphosate. The second form is EPSP II, which is found in glyphosate-resistant bacteria and is not inhibited by glyphosate. It is thus the gene encoding an EPSPS II that has been used to genetically engineer crops to confer resistance/tolerance to this particular herbicide [34].

Environmental Impact of Transgenic Crops

The adoption of transgenic crops, and in particular herbicide-tolerant crops, has resulted in significant changes in agronomic practices. Many clear benefits to both the grower and consumer can be identified with transgenic crops that are currently grown, not least in the often significant increases in productivity accompanied by reductions in synthetic pesticide usage (Tables 1 [6] and 2 [7–20, 35–37]), thus making transgenic crops more environmentally sustainable. Despite the fact that such crops have now been commercialized for some 15 years, alongside these benefits, concerns are still being expressed as to their safety, particularly in terms of their impact on human health and the environment. It is imperative that these concerns are addressed in a scientifically sound and appropriate manner. Thus, assessing the environmental consequences of transgenic crops is an important prerequisite to their commercialization.

In many regions of the world, regulatory frameworks are in place to ensure that all pre-commercial transgenic crops are evaluated for potential impacts on human health, animal health, and the environment according to established standards of risk assessment and current scientific knowledge, prior to authorizations for import or planting being granted. The environmental risk assessment for such crops follows the same fundamental principles as other risk assessment

schemes, i.e., risk is a function of hazard and exposure. However, one of the main differences that sets risk assessment of transgenic crops apart from other risk assessment procedures is that it is highly dependent on the crop and the introduced trait; hence, a case-by-case approach is required [38].

Impact of Insect-Resistant Transgenic Crops on Beneficial Insects

For any technology to be acceptable to the public at large, the perceived benefits have to outweigh any potential/perceived risk – this requirement is equally true for transgenic crops [39]. Common to all transgenic crops currently grown, general environmental concerns relate to the potential for horizontal gene transfer, gene flow, and invasiveness to occur both in managed and natural environments. However, in the case of insect-resistant transgenic crops, perhaps, the greatest concerns relate to their non-target effects and, in particular, their effects on beneficial insects such as natural enemies (predators and parasitoids) and pollinators.

An important consideration is the likelihood of exposure of the transgene product to the non-target insect, and indeed the different routes of exposure. The most obvious exposure route for non-target herbivores is through direct ingestion of plant material, although this will be influenced by the mode of insect feeding and spatial and temporal expression patterns of the transgene product within the plant. While exposure to pollinators is via the nectar and pollen, exposure routes to natural enemies are more diverse. Although many predators and parasitoids, particularly in the adult stage, are facultative herbivores and thus can be exposed to transgene products directly from consuming plant tissues (pollen, nectar), the larval stages are more likely to be exposed from consuming insects that have themselves fed on plant tissues where the transgene product has been expressed and accumulated. It is pertinent to point out that in some cases, particularly when the host is a sap-sucking insect such as aphids, that the natural enemy is rarely exposed to the transgene product; however, the scenario is quite different when the host is a chewing insect, in which case the likelihood of the natural enemy being exposed is very high.

Impact on Natural Enemies Agro-ecosystems consist of complex trophic interactions, with many aspects of the physiology, ecology, and behavior of the organisms present being governed by interactions with organisms from the same or different trophic levels. Since plants are the basis of these food webs on which all organisms at higher trophic levels depend, and since herbivorous insect species at the second trophic level have an important position in food webs, together with the fact that approximately 50% of all insect species are plant feeders, the potential for exposure to non-targets, either directly (bi-trophic) or indirectly (tri-trophic), is high. This is particularly true for predators and parasitoids which play an important role in suppressing insect pest populations both in the field and under specialized cultivation systems. Because of this important role and because expression of transgenes that confer enhanced levels of resistance to insect pests is of particular relevance (since they are aimed at manipulating the biology of organisms in a different trophic level to that of the plant), these beneficial insects have been the focus of numerous major studies to evaluate the non-target effects of transgenic crops. The nature of these investigations has been very varied ranging from detailed studies at the molecular/biochemical level under controlled environmental conditions, to glass-house trials using deliberately infested plants and released natural enemies, to effects at the population level in the field. Further, they have involved delivering the transgene product both in artificial diet and *in planta* and either directly (at the bi-trophic level) or via a tri-trophic (plant-pest-natural enemy) interaction. Recent studies relating to both the effects of *Bt*-expressing crops and crops expressing molecules still under development are summarized in [Tables 3 \[40–72\]](#) and [4 \[50, 51, 73–102\]](#) for predators and parasitoids, respectively.

As mentioned above, prior to their commercialization, all transgenic crops have to undergo rigorous risk assessment. However, despite a global commodities market, the risk assessment process itself varies from region to region, as does the non-target organisms selected for testing. Obviously, the latter will vary depending upon geographical and climatic regions; however, there are concerns that there should be greater uniformity in the species selected; these should at least be representative of the major ecosystem functions. In

an attempt to address this short-coming, Romeis et al. [\[103\]](#) have recently refined the 3-tier risk assessment to evaluate potential adverse impacts of insect-resistant transgenic crops on non-target arthropods, with specific recommendations for early-tier laboratory studies used in the risk assessment process. Although this risk assessment protocol has been primarily developed for crops expressing *Bt* toxins, the concepts put forward apply to other arthropod-active proteins. Typically, the risk assessment follows a tiered approach that starts with laboratory studies under worst-case exposure conditions; such studies have a high ability to detect adverse effects on non-target species, if present. Clear guidance on how such data are produced in laboratory studies assists the product developer and risk assessors. The need for a high level of reproducibility and clearly defined risk hypotheses contribute to the robustness of, and confidence in, the environmental risk assessments of transgenic plants. Further, these authors emphasized that confidence in the results of early-tier laboratory studies is a precondition for the acceptance of data across regulatory jurisdictions and should encourage agencies to share useful information and thus avoid redundant testing. When evaluating potential risks of a given technology, it is important to use relevant comparators. This holds true for risk assessment of transgenic crops. Few studies have actually been designed to directly compare recombinant DNA technology with conventional pest control strategies, although recent studies by Mulligan et al. [\[68, 69\]](#) directly compared the non-target effects of the synthetic pesticide cypermethrin with oilseed rape plants expressing a cysteine protease inhibitor against two predators. While neither form of pest control treatment negatively affected the carabid, the effects of the transgenic crop on the lacewing were significantly lower than with the commonly used pesticide.

Recent studies have performed meta-analyses on a modified public database [\[104\]](#) to synthesize current knowledge about the effects of *Bt* cotton, maize, and potato on the abundance and interactions of arthropod non-target functional guilds [\[105\]](#). Overall, they show no uniform effects of *Bt* cotton, maize, and potato on the functional guilds of non-target arthropods. In fact, use of and type of insecticides influenced the magnitude and direction of effects; and insecticide effects were much larger than those of *Bt* crops. Similarly,

Transgenic Crops, Environmental Impact. Table 3 Impacts of transgenic crops and transgene products on predators

Protein [1]	Transgenic plant or diet	Pest	Natural enemy	Effects on natural enemy	Reference
Bt	Corn (Cry1Ab)	Direct feeding (pollen)	Several (Coleoptera, Heteroptera, Neuroptera)	No effects on predators in both laboratory and field experiments	[40]
	Corn (Cry3Bb1)	<i>Diabrotica</i> spp. (Col: Chrysomelidae)	Several (Araneae, Carabidae, Staphylinidae)	No consistent negative effect	[41]
	Corn (Cry1Ab)	<i>Ostrinia nubilalis</i> ; <i>Spodoptera littoralis</i> (Lep: Noctuidae)	<i>Caladenia carnea</i>	Bt-fed prey increased predator mortality and development times	[42]
		<i>S. littoralis</i>	<i>C. carnea</i>	Negative effects, as observed previously, determined to be due to reduced prey quality	[43]
		<i>Tetranychus urticae</i> (Acari)	<i>Stethorus punctillum</i> (Col: Coccinellidae)	No effects in both laboratory and field experiments	[44]
		Direct feeding (pollen)	Spiders (Araneae)		[45, 46]
	Corn (Cry3Bb1)	<i>Rhopalosiphum maidis</i> (Hom: Aphididae)	<i>Coleomegilla maculate</i> (Col.: Coccinellidae)	No effects when fed non-target aphid prey	[47]
	Corn (VIP3A + Cry1Ab)	Lepidoptera pests	Several (13 arthropod orders)	Large-scale study showed no negative effects of stacked traits over conventional corn	[48]
	Cotton (Cy1Ac)	<i>Aphis gossypii</i> (Hem. Aphididae)	<i>Chrysopa pallens</i> (Neu: Chrysopidae)	No effect	[49]
	Cotton (Cry1Ac)	Several	Predators (several)	Minor reductions in predator density in the field	[50, 51]
	Cotton (Cry1Ac/ Cry2Ab)	Lepidopterous pests		Predator numbers similar or higher in Bt cotton field plots	[52,53]
	Cotton (Cry1Ac)	<i>Spodoptera exigua</i> , <i>Helicoverpa zea</i> (Lep: Noctuidae)	<i>Geocoris punctipes</i> (Het: Lygaeidae)	No effects in field experiments	[54]
		<i>S. exigua</i>	<i>Podisus maculiventris</i>	No effects	[55]
		Lepidopterous pests	<i>C. carnea</i> ; <i>Orius tristicolor</i> (Het: Anthocoridae)	No effects in field experiments	[56]
	Potato (Cry3Aa)	<i>Leptinotarsa decemlineata</i>	Several heteropteran predators and spiders	No effect of predator densities in the field	[57]

Transgenic Crops, Environmental Impact. Table 3 (Continued)

Protein [1]	Transgenic plant or diet	Pest	Natural enemy	Effects on natural enemy	Reference
	Potato (Cry3Aa)	<i>L. decemlineata</i>	Coleoptera, Araneae	No effects on field pitfall trap capture numbers	[58]
	Potato (Cry3)	<i>Myzus persicae</i>	<i>Hippodamia convergens</i> (Col: Coccinellidae)	No effect	[59]
	Potato (Cry3A)	Lep. Hem.	Several (Heteroptera)	No effects on development time	[60]
	Potato		<i>Harmonia axyridis</i> , <i>Nebria brevicollis</i>	No effects	[61]
CpTI	Injected prey	<i>L. oleracea</i>	<i>P. maculiventris</i> (Het: Pentatomidae)	Reduced growth of predators	[62]
	Potato		<i>P. maculiventris</i>	No effects	[62]
	Strawberry	<i>Otiorhynchus sulcatus</i> (Col: Curculionidae)	Carabids and others	Field abundance not affected	[63]
HvCPI-1 C68 → G	Potato	<i>L. decemlineata</i>	<i>P. maculiventris</i>	No effect	[64]
		<i>S. littoralis</i>			[64]
MTI-2	Oilseed rape	<i>Plutella xylostella</i>	<i>Pterostichus madidus</i> (Col: Carabidae)	No effects on reproductive fitness; female weight gain reduced at first but compensated for later	[65]
Aprotinin (bovine pancreatic or bovine spleen trypsin inhibitor) (BPTI/BSTI)		<i>Helicoverpa armigera</i>	<i>Harpalus affinis</i> (Col: Carabidae)	Beetles consumed less prey after 24 h of exposure to inhibitor-fed prey	[66]
BPTI/BSTI		<i>H. armigera</i>	<i>Nebria brevicollis</i> (Col: Carabidae)	Transient minor changes in adult beetle weights	[67]
OCI		Direct feeding	<i>C. carnea</i>	No effect via contaminated pollen	[68]
		Oilseed rape	<i>Deroceras reticulatum</i> (Mollusca)	<i>Pterostichus melanarius</i> (Col. Carabidae)	No effects on beetle mortality, weight gain, or food consumption
	Oilseed rape	<i>P. xylostella</i> (Lep: Plutellidae)	<i>H. axyridis</i>	No effect on survival or development of ladybird	[70]
	Potato	<i>L. decemlineata</i> (Colorado potato beetle)	<i>Perillus bioculatus</i> (Het: Pentatomidae)	No effects	[71]
Soybean trypsin inhibitor (SBTI)	Diet	Direct feeding	<i>C. carnea</i>		[72]

Transgenic Crops, Environmental Impact. Table 4 Impacts of transgenic crops and transgene products on parasitoids

Protein [1]	Transgenic plant or diet	Pest	Natural enemy	Effects on natural enemy	Reference
Bt	Corn (Cry1Ab)	<i>Chilo partellus</i> (Lep: Crambidae)	<i>Diaeretiella rapae</i> (Hym: Braconidae)	Reduced survival due to host mortality, smaller cocoons and adults	[73]
	Corn (Cry1Ab) (field)	<i>Ostrinia nubilalis</i> (Lep: Crambidae)	<i>Macrocentrus cingulum</i> (Hym: Braconidae)	Reductions of 29–60% is numbers of wasps found in the field	[74]
	Corn (Cry1Ab)	<i>Spodoptera frugiperda</i> (Lep: Noctuidae)	<i>Campoletis sonorensis</i> (Hym: Ichneumonidae)	Wasps were significantly smaller when developing in Bt-fed hosts	[75]
		<i>Eoreuma loftini</i> (Lep: Pyralidae)	<i>Parallorhogas pyralophagus</i> (Hym: Braconidae)	Various aspects of parasitoid biology (not all) negatively affected	[76]
	Cotton (Cry1Ac)	<i>Helicoverpa armigera</i> (Lep: Noctuidae)	<i>Microplitis mediator</i> (Hym: Braconidae)	Wasp survival and development negatively affected	[77, 78]
		<i>Pseudoplusia includens</i> (Lep: Noctuidae)	<i>Cotesia marginiventris</i> (Hym: Braconidae), <i>Copidosoma floridanum</i> (Hym: Encyrtidae)	Development times for both wasp species negatively affected. Adult longevity for <i>C. marginiventris</i> reduced	[79]
	Cotton (Cry1Ac) (field)	Hemipteran pests	Aphelinid parasitoids	Small reduction of parasitoid population density relative to non-Bt cotton	[50, 51]
	Diet (Cry1Ac)	<i>Spodoptera litura</i> /H. <i>armigera</i> (Lep: Noctuidae)	<i>Meteorus pulchricornis</i> (Hym: braconidae); <i>Cotesia kazak</i> (Hym: Braconidae)	Survival of both unaffected in Bt-fed <i>S. litura</i> . <i>M. pulchricornis</i> negatively affected by Bt –fed <i>H. armigera</i>	[80]
	Oilseed rape (Cry1Ac)	<i>Plutella xylostella</i> (Lep: Plutellidae)	<i>Cotesia plutellae</i> (Hym: Braconidae)	No effect when Bt-resistant hosts were parasitized	[81]
	Broccoli	<i>P. xylostella</i>	<i>Diadegma insulare</i> (Hym: Ichneumonidae)	No effect on parasitoid when exposed by Bt-resistant hosts	[82]
	Pine (Cry1Ac)	<i>Pseudocoremia suavis</i> (Lep: Geometridae)	<i>M. pulchricornis</i>	No effect on parasitoid developmental parameters	[83]
	Diet (Cry9Aa)	<i>Galleria mellonella</i> (Lep: Gelechiidae)	<i>Exorista larvarum</i> (Dip: Tachinidae)	No effect on fly parasitoid when exposed to factitious Bt-fed hosts	[84]
	Tobacco (Cry1Ab) (field)	<i>Heliothis virescens</i> (Lep. Noctuidae)	<i>C. sonorensis</i>	Rates of parasitism increased	[85]
CpTI	Diet or potato	<i>Lacanobia oleracea</i> (Lep: Noctuidae)	<i>Eulophus pennicornis</i> (Hym: Eulophidae)	Fewer hosts parasitized, no effects on parasitoids	[86]

Transgenic Crops, Environmental Impact. Table 4 (Continued)

Protein [1]	Transgenic plant or diet	Pest	Natural enemy	Effects on natural enemy	Reference	
CpTI	Diet	Direct feeding (adult wasps)	<i>E. pennicornis</i>	No effects	[87]	
CpTI and Bt	Cotton	<i>H. armigera</i>	<i>Microplitis mediator</i> (Hym: Braconidae)	No greater effects than with Bt-cotton	[78]	
	Cotton pollen	Direct feeding	<i>Trichogramma chilonis</i> (Hym: Trichogrammatidae)	No adverse effects due to CpTI	[88]	
OC1	Diet	<i>Macrosiphum euphorbiae</i> (Hom: Aphididae)	<i>Aphelinus abdominalis</i> (Hym: Braconidae)	Parasitoid fitness impaired	[89]	
	Potato	<i>M. euphorbiae</i>	<i>Aphidius nigripes</i> (Hym: Braconidae)	Wasp size and fecundity increased on OC1 line	[90]	
	Oilseed rape	<i>Myzus persicae</i> (Hom: Aphididae)	<i>Diaeretiella rapae</i> (Hym: Braconidae)	No consistent effects on adult wasp emergence and sex ratio; no effects on control of aphids	[91]	
OC1 IAD86 (for nematode control)	Potato (field)	<i>M. euphorbiae</i> , <i>M. persicae</i>	<i>Aphidius ervi</i> (Hym: Braconidae)	No effects on percent parasitism, adult wasp emergence, parasitoid communities more diverse on GM plants	[92]	
Con A	Diet	Direct feeding	<i>E. pennicornis</i>	Higher doses reduced adult longevity and reproductive fitness	[87], [93]	
GNA		<i>M. euphorbiae</i>	<i>A. abdominalis</i>	No effect via host feeding on aphids Reduced size and longevity of adults when exposed via host	[94, 95]	
		Direct feeding (adult wasps)	<i>Aphidius colemani</i> (Hym: Braconidae)	Higher doses reduced adult longevity	[96]	
		Honeydew	Direct feeding (adult wasps)	<i>A. ervi</i>	Indirect negative affect potentially due to altered honeydew composition	[97]
		Potato/diet	<i>M. persicae</i>	<i>A. ervi</i>	No effects via potato, dose-dependent effects on development via diet	[94, 95]
		Sucrose diet	Direct feeding	<i>Cotesia glomerata</i> (Hym: Braconidae)	Higher doses reduced longevity	[96]
		Diet/ Sugarcane	<i>Diatraea saccharalis</i> (sugarcane borer)	<i>Cotesia flavipes</i> (Hym: Braconidae)	Small negative effects on parasitism. No effects on host location of prey or parasitism	[98, 99]
		Potato/Diet	<i>L. oleracea</i>	<i>E. pennicornis</i>	No effect on parasitism success	[100]
		Diet	Direct feeding		Reduced adult longevity and reproductive fitness	[87], [93]

Transgenic Crops, Environmental Impact. Table 4 (Continued)

Protein [1]	Transgenic plant or diet	Pest	Natural enemy	Effects on natural enemy	Reference
	Host diet/host injection	<i>L. oleracea</i>	<i>E. pennicornis</i>	Negative effects on the survival of parasitoid larvae	[93]
	Sugarcane	<i>E. loftini</i>	<i>P. pyralophagus</i>	Reduced size and longevity of adult wasps	[101]
	Tomato/potato/diet	<i>L. oleracea</i>	<i>Meteorus gyrator</i> (Hym: Braconidae)	No effects	[102]
	Sucrose diet	Direct feeding (adult wasps)	<i>Trichogramma brassicae</i> (Hym: Trichogrammatidae)	Antifeedant; high dose reduced longevity	[96]

Duan et al. [106] performed meta-analyses comparing results for non-target invertebrates exposed to *Bacillus thuringiensis* (Bt) Cry proteins in laboratory studies with results derived from independent field studies examining effects on the abundance of non-target invertebrates. In this case, the findings support the assumption that laboratory studies of transgenic insecticidal crops show effects that are either consistent with, or more conservative than, those found in field studies.

In combination with robust field data [107], the evidence strongly supports Bt crops as an environmentally neutral technology, especially in comparison to insecticides. These findings are thus in agreement with those of Mulligan et al. [68, 69].

Impact on Pollinators Many of the world's crops depend on insects for pollination and it is critically important that agricultural biotechnology does not disrupt this essential "ecosystem service" – as such, pollinators are critical to agriculture, in addition to their vital role in helping maintain biodiversity. Thus, representative pollinators form an important part of the risk assessment process and much of what has been stated for natural enemies applies to these species too. In common with natural enemies, transgenic plants can impact on pollinators in two ways, either directly (via the transgenic plant itself posing a hazard to the pollinator), or indirectly (via the use of a transgenic crop affecting other ecological requirements of the pollinator). Of the extensive field trials carried out to date with either Bt-expressing crops, or HT transgenic crops, no

deleterious effects on pollinators have been reported [108]. Furthermore, results from the many studies carried out to date on transgenic plants still under development suggest that only a few of the lectins and protease inhibitors tested could present dose-dependent hazards to bees, and then only if expressed in the pollen at sufficiently high levels; most current promoters used, albeit constitutive promoters, only express at very low levels, if at all, in the pollen.

In addition to evaluating the potential impacts of transgenic crops on pollinator species, it is also important to address issues relating to the ability of these species to collect and transport pollen between transgenic and non-transgenic crops, i.e., their role in pollen dispersal and hence gene flow. Recombinant proteins are usually expressed at low levels in the pollen of transgenic plants due to the use of specific "constitutive" promoters in the gene constructs. However, the transgenes themselves are present in DNA contained in the pollen, and therefore may be transferred to other related plants via the activities of the pollinator species.

Impact of Herbicide-Tolerant Transgenic Crops

The wide-scale growing of glyphosate-tolerant crops has resulted in an increase in glyphosate usage at the expense of other herbicides [4, 109–111]. However, despite this increase in glyphosate use, overall, the use of HT crops has resulted, overall, in a significant reduction in pesticides. This has been accompanied by a reduction in tillage which has an additional benefit

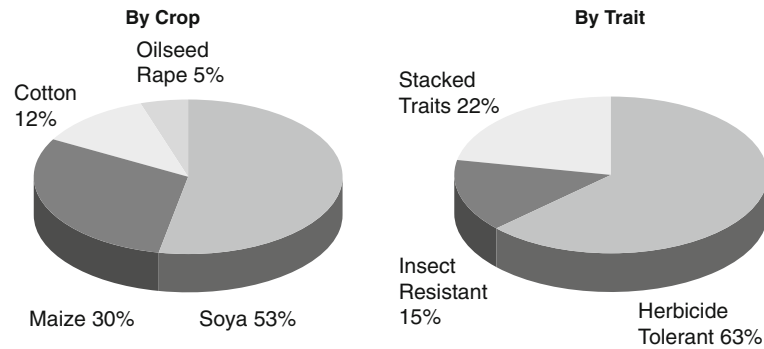
of reducing the use of petroleum-based fuels as well as an implicit gain in time-use efficiency by growers [112]. Furthermore, HT crops have dramatically changed the crop cultivars selected by growers and has hastened the development of new transgenic crops for commercial distribution worldwide [32, 113].

Weed control based on HT crops is efficient and cost effective and provides considerable savings in terms of both time and labor [114]. However, despite this and other obvious benefits of the technology (see above), there is much opposition in some quarters to these crops – possibly more so than for insect-resistant transgenic crops. One of the reasons for this opposition stems from a lack of effective communication and understanding between the general public and the scientific and agricultural communities [21, 115]. Designing a risk assessment to test specific questions to also address societal concerns may help address this issue since it would provide an opportunity for society to participate in the regulatory decisions affecting HT crops [116]. It is possible that a major part of the concern expressed about HT crops is that the technology is associated with herbicide use, which in itself is perceived as risky by the public sector [117]. The mode of action of glyphosate is well understood and in fact was one of the first commercially successful herbicides to have an identified enzymatic site of action in plants. The selective and specific interaction of glyphosate with EPSPS accounts for its potent herbicidal properties and low toxicity to other life forms; as such, this will significantly reduce the potential for direct non-target effects and should negate many of these concerns.

Despite their undoubted success and major contribution to global agriculture, potential risks associated with the technology must be assessed, not least in terms of the increased use of glyphosate [112, 118–121]. However, this apart, the primary environmental risk associated with HT crops relates to their impact on weed population shifts, whether expressed as the rise in economic prominence of a new weedy species or the evolution of glyphosate-tolerant weed biotypes. This position is supported by the herbicide resistance risk analysis which suggests that glyphosate-based systems are at high risk of selecting for glyphosate-tolerant weeds [122], although not all research supports this view point [123].

Impact on Biodiversity Agriculture itself has had a major impact on biodiversity, often manifested in a marked decline in the abundance of species [124–126]. While it is suggested that the ecosystem effects of HT crops have been minimal [127], the indirect impact of these crops on the agro-ecosystem, particularly as a result of changes in tillage and weed management tactics, is important [128]. There is conflicting evidence in the literature with some reports suggesting an increase in species diversity in HT-based systems while other reports suggest a reduction [129–133]. These apparent contradictions may be due to the fact that effects are often specific to the crop in question and dependent on the different weed management tactics used for HT crops compared to non-transgenic crops [134, 135]. For example, deployment of HT crops was shown to have a negative impact on butterfly population densities as an indirect effect of good weed control, reflecting a lack of nectar availability [136]. In contrast, little effect was observed for bees, gastropods, and other invertebrates. Furthermore, the effects of such crops on the soil biota were found to be negligible [132]. Thus, there appears to be both favorable and unfavorable data on the effects of HT crops on biological diversity [32]. The critical consideration is that these effects are highly dependent on specific crop and management tactics. It is likely that any unfavorable effect on biological diversity could be ameliorated by subtle manipulation of the HT-based system.

Gene Flow Another pervasive problem with HT crops is their coexistence with non-GM crops. Three important considerations have to be addressed: (1) introgression of the trait via pollen (pollen drift); (2) containment of plant products during the production year (grain segregation); and (3) volunteer HT plants in following years [115]. While HT crops and their non-transgenic counterparts can, and do, coexist, and while grain segregation and controlling volunteers is feasible [115, 137], controlling/preventing introgression of the HT trait via pollen movement in open-pollinated crops such as maize is considerably more difficult [138–141]. A number of factors affect the success of maize pollen movement and subsequent pollination, and generally, the greater the distance between the pollen source and donor, the less likely is the introgression of the transgenic trait [139, 140].



Transgenic Crops, Environmental Impact. Figure 2
Global status of biotech crops for 2008

Given the tolerance levels established for some transgenic traits in non-transgenic crops, the isolation distances required to mitigate the risks of gene flow may be too great to be realistic in commercial maize production systems [142]. Other open-pollinated crops have also been scrutinized with significant legal ramifications [143–146]. However, because of the possibility of “contamination” as a consequence of gene flow, many countries require buffer zones between fields growing transgenic crops and their non-transgenic counterparts; the size of these buffer zones will vary, depending upon a number of different factors including the crop itself. It is suggested that the issues of the coexistence of such crops with non-transgenic crops will continue to be a concern as long as there are economic differences between the crop cultivar types [147–149].

Future Directions

The major traits currently commercialized predominantly confer resistance to biotic stress, with herbicide-tolerant (HT) crops occupying the market share (Fig. 2). In 2009, SmartStax, a novel biotech maize containing eight different genes for insect and herbicide resistance, was approved for commercialization in the USA. These so-called first-generation transgenic crops, expressing improved agricultural traits are often perceived as being of benefit to the grower rather than to the consumer *per se*. Although the seed is more expensive, such crops lower the costs of production by reducing inputs of machinery, fuel, and chemical

pesticides. In addition, due to more effective pest control, crop yields are often higher (Tables 1 [6] and 2 [7–20]). However, it is important to appreciate the environmental and health benefits of growing these first-generation transgenic crops, many of which are associated with reduced spraying of highly toxic chemical insecticides and herbicides. In terms of environmental benefits, these include controlling farm runoff that otherwise pollutes water systems and reduced mechanical weeding, so reducing loss of topsoil [150], while the major health benefits are a consequence of reduced pesticide exposure for farmers and rural laborers, and lower pesticide residues for consumers.

In addition to transgenic crops with increased tolerance to biotic stress, drought-tolerant maize is expected to be deployed in the USA in 2012 and sub-Saharan Africa in 2017. Other transgenic crops on the horizon include adoption of Golden Rice by the Philippines in 2012 and Bangladesh and India before 2015. Other smaller hectareage crops are also expected to be approved by 2015, including potatoes with pest and/or disease resistance, sugarcane with quality and agronomic traits, and disease-resistant bananas [3]. Interestingly, wheat remains the last major staple crop without approved biotech traits. However, political will for the crop is growing globally and many authorities suggest that China may be the first country to approve biotech wheat as early as 5 years from now.

As the science develops, so does the technology, but irrespective of which particular generation of transgenic crops is being considered, their environmental impact is of prime importance. Any risk assessment

process must thus keep pace with the changing technology and the development of novel crops expressing novel traits.

Conclusion

Although one of the major concerns of recombinant DNA technology relates to its impact on non-target organisms, and thus on biodiversity, these fears have not, in the main, been realized, although there have been some well-publicized cases to the contrary. Concern over the potential for *Bt*-expressing maize to have negative effects on the Monarch Butterfly (*Danaus plexippus*) population was voiced following the publication of lab-based studies that demonstrated that unrealistically high levels of pollen from such plants had a deleterious effect [151]. However, subsequent large-scale field trials demonstrated this not to be the case, one factor being that when the maize was in flower, the Monarchs were not present [152–157]. Thus, in this instance, while the potential hazard was high, exposure was negligible resulting in effectively zero risk. Other examples include reports of initial studies concerning toxicity of *Bt* maize–fed hosts toward the predator *Chrysoperla carnea* (the green lacewing) via a tri-trophic interaction [42]. However, subsequent studies demonstrated that *Bt* Cry1Ab was not toxic to the larvae but that the effects reported were mediated by prey quality [158]. Studies of this type emphasize the need not only to place them within an ecological context, but also to use appropriate comparators in the risk assessment process. Studies such as those carried out by Hilbeck [42] and colleagues emphasize the importance of scientific rigor and the need for demonstration of “cause and effect.”

While numerous studies have now been carried out to evaluate the environmental safety of transgenic crops on beneficial insects such as natural enemies that play an important role in biological control and pollinators, in the vast majority of cases, few negative effects have actually been demonstrated. Interestingly, in the majority of cases studied to date regarding natural enemies, it is apparent that the predator/parasitoid is often able to avoid the toxic effects of the different insecticidal proteins being expressed, despite exposure at physiologically relevant levels. Further, there is evidence that the transgene product is diluted as it passes

through the different trophic levels. For pollinators, this lack of toxicity is attributed to lack of receptors, in the case of *Bt*, and lack of exposure when expressed in the transgenic plant.

In addition to their potential impact on biodiversity, other environmental concerns relating to the deployment of transgenic crops are the potential for gene flow, particularly in the case of HT crops. However, the potential for gene flow is highly dependent upon the plant species [159]. Gene flow in crops can occur via pollen and via seed, the latter potentially affecting agriculture temporally and on a much larger scale than gene flow attributable to pollen [160]; in general, it is acknowledged that the risks of unintended trait movement are difficult to assess [161]. It is, however, important to point out that gene flow is no different in transgenic crops than in non-transgenic cultivars and that gene flow from transgenic crops is a reality [162]. To expect compliance of zero gene flow is neither reasonable nor realistic, and an acceptable tolerance must be established for the coexistence of transgenic crops with their non-transgenic counterparts.

Bibliography

1. Royal Society Policy Document 11/09 (2009) Reaping the benefits: science and the sustainable intensification of global agriculture. Royal Society, London
2. FAO (2000) Agriculture: Towards 2015/30. Technical interim report, April. Economic and social department. Rome, Italy
3. James C (2008) Global status of commercialized biotech/GM crops: 2008. The First Thirteen Years, 1996 to 2008. ISAAA Brief 39-2008: Executive Summary. ISAAA, New York
4. Service RF (2007b) A growing threat down on the farm. *Science* 316:1114–1117
5. Meyer B (2008) Biotech crops gaining favor around the globe. Learfield Communications, Inc., Jefferson City
6. Zilberman D, Sexton SE, Marra M, Fernandez-Cornejo J (2010) The economic impact of genetically engineered crops. <http://www.choicesmagazine.org/>
7. Delannay X, Bauman TT, Beighley DH, Buettner MJ, Coble HD, DeFelice MS, Derting CW, Diedrick TJ, Griffin JL, Hagood ES, Hancock FG, Hart SE, LaVallee BJ, Loux MM, Lueschen WE, Matson KW, Moots CK, Murdock E, Nickell AD, Owen MDK, Paschall EH II, Prochaska LM, Raymond PJ, Reynolds DB, Rhodes WK, Roeth FW, Sprankle PL, Tarochionc LJ, Tinius CN, Walker RH, Wax LM, Weigelt HD, Padgett SR (1995) Yield evaluation of a glyphosate-tolerant soybean line after treatment with glyphosate. *Crop Sci* 35:1461–1467

8. Roberts RK, Pendergrass R, Hayes RM (1998) Farm-level economic analysis of roundup Ready™ soybeans. Paper presented at the Southern Agricultural Economics Association meeting, Little Rock, 1–4 Feb 1998
9. Arnold JC, Shaw DR, Medlin CR (1998) Roundup Ready™ programs versus conventional programs: efficacy, varietal performance, and economics. *Proc South Weed Sci Soc* 51: 272–273
10. Marra M, Carlson G, Hubbell B (1998) Economic impacts of the first crop biotechnologies. Available at <http://www.ag.econ.ncsu.edu/faculty/marra/firstcrop/imp001.gif>
11. Fernandez-Cornejo J, Klotz-Ingram C, Jans S (2002) Farm-level effects of adopting herbicide-tolerant soybeans in the USA. *J Agric Appl Econ* 34:149–163
12. Duffy M (2001) Who benefits from biotechnology? Paper presented at the American Seed Trade Association meeting, Chicago, 5–7 Dec 2001
13. Marra MC (2004) Review of the economic and environmental impacts of agbiotech: a global perspective. *AgBioForum* 6:144–145, Nicholas GK (ed)
14. Bernard JC, Pesek JD Jr, Fan C C (2004) Delaware farmers adoption of GE soybeans in a time of uncertain U.S. adoption. *Agribus Int J* 20:81–94
15. Qaim M, Traxler G (2005) Roundup Ready soybean in Argentina: farm level and aggregate welfare effects. *Agric Econ* 32:73–86
16. Vencill WK (1996) Weed management systems utilizing herbicide-resistant cotton. *Proceedings of the Beltwide Cotton Conferences Cotton Weed Science Research Conference* 2:1532–1533
17. Keeling W, Dotray P, Jones C, Sunderland S (1996) Roundup™ ready cotton: a potential new weed management tool for the Texas high plains. *Proceedings of the Beltwide Cotton Conferences Cotton Weed Science Research Conference* 2:1529
18. Goldman IL (1998) From out of old fields comes all this new corn: An historical perspective on heterosis in plant improvement. *Concepts and breeding of heterosis in crop plants*. Book series: CSSA Special Publications Issue: 25, pp 1–12
19. Culpepper AS, York AC (1998) Weed management in glyphosate-tolerant cotton. *J Cotton Sci* 4:174–185
20. Fernandez-Cornejo J, Klotz-Ingram C, Jans S (2000) Farm-level effects of adopting genetically engineered crops in the USA. In: Lesser WL (ed) *Transitions in agrobiotech: economics of strategy and policy*. Food Marketing Research Center, University of Connecticut and Department of Research Economics, University of Massachusetts, pp 57–72
21. Kraye von Krauss MP, Casman EA, Small MJ (2004) Elicitation of expert judgments of uncertainty in the risk assessment of herbicide-tolerant oilseed crops. *Risk Anal* 24:1515–1527
22. de Maagd RA, Bravo A, Crickmore N (2001) How *Bacillus thuringiensis* has evolved specific toxins to colonize the insect world. *Trends Genet* 17:193–199
23. Song XX, Wang SM (2001) Status and evaluation on the expression of cotton varieties in the production in China in the past 20 years. *Cotton Sci* 13:315–320
24. He KL, Wang ZY, Zhang YJ (2009) Monitoring *Bt* resistance in the field: China as a case study. In: Ferry N, Gatehouse AMR (eds) *Environmental impact of genetically modified crops*. Wallingford, CAB International, pp 344–359, Chap 16
25. Tabashnik BE, Carrière Y (2009) Insect resistance to genetically modified crops. In: Ferry N, Gatehouse AMR (eds) *Environmental impact of genetically modified crops*. CAB International, Wallingford, pp 74–100, Chap 5
26. Christou P, Capell T, Kohli A, Gatehouse JA, Gatehouse AMR (2006) Recent developments and future prospects in insect pest control in transgenic crops. *Trends Plant Sci* 11:302–308
27. Gatehouse JA (2002) Plant resistance towards insect herbivores: a dynamic interaction. *New Phytol* 156:145–169
28. Gatehouse JA (2008) Biotechnological prospects for engineering insect-resistant plants. *Plant Physiol* 146: 881–887
29. Malone LA, Gatehouse AMR, Barratt BIP (2008) Beyond Bt: alternative strategies for insect-resistant crops. In: Romeis J, Shelton T, Kennedy G (eds) *Integration of insect-resistant genetically modified crops within integrated pest management programs*, Series on Progress in Biological Control. Springer, New York
30. Price DR, Gatehouse JA (2008) RNAi-mediated crop protection against insects. *Trends Biotechnol* 26:393–400
31. James C (2009) Global status of commercialized biotech/GM crops: 2009. ISAAA Brief No. 41. ISAAA: Ithaca
32. Duke SO (2005) Taking stock of herbicide-resistant crops ten years after introduction. *Pest Manag Sci* 61:211–218
33. Duke SO, Powles SB (2008) Glyphosate: a once-in-a-century herbicide. *Pest Manag Sci* 64:319–325
34. Mazur S, Falco C (1989) The development of herbicide resistant crops. *Annu Rev Plant Physiol Plant Mol Biol* 40:441–470
35. Betz FS, Hammond BG, Fuchs RL (2000) Safety and advantages of *Bacillus thuringiensis*-protected plants to control insect pests. *Regul Toxicol Pharmacol* 32:156–173
36. Cattaneo MG, Yafuso C, Schmidt C, Huang CY, Rahman M, Olson C, Eilers-Kirk C, Orr BJ, Marsh SE, Antilla L, Dutilleul P, Carrière Y (2006) Farm-scale evaluation of the impacts of transgenic cotton on biodiversity, pesticide use, and yield. *Proc Natl Acad Sci USA* 103:7571–7576
37. Thies JE, Devare MH (2007) An ecological assessment of transgenic crops. *J Dev Stud* 43:97–129
38. Tencalla FG, Nickson TE, Garcia-Alonso M (2009) Environmental risk assessment. In: Ferry N, Gatehouse AMR (eds) *Environmental impact of genetically modified crops*. CAB International, Wallingford, pp 61–73, Chap 4
39. Waltz E (2009) Battlefield. *Nature* 461:27–32
40. Pilcher CD, Obrycki JJ, Rice ME, Lewis LC (1997) Preimaginal development, survival, and field abundance of insect predators on transgenic *Bacillus thuringiensis* corn. *Environ Entomol* 26:446–454

41. Bhatti MA, Duan J, Head G, Jiang CJ, McKee MJ, Nickson TE, Pilcher CL, Pilcher CD (2005) Field evaluation of the impact of corn rootworm (Coleoptera: Chrysomelidae)-protected Bt corn on ground-dwelling invertebrates. *Environ Entomol* 34:1325–1335
42. Hilbeck A, Baumgartner M, Fried PM, Bigler F (1998) Effects of transgenic *Bacillus thuringiensis* corn-fed prey on mortality and development time of immature *Chrysoperla carnea* (Neuroptera: Chrysopidae). *Environ Entomol* 27: 480–487
43. Dutton A, Klein H, Romeis J, Bigler F (2002) Uptake of Bt-toxin by herbivores feeding on transgenic maize and consequences for the predator *Chrysoperla carnea*. *Ecol Entomol* 27:441–447
44. Alvarez-Alfageme F, Ferry N, Castanera P, Ortego F, Gatehouse AMR (2008) Prey mediated effects of Bt maize on fitness and digestive physiology of the red spider mite predator *Stethorus punctillum* Weise (Coleoptera: Coccinellidae). *Transgenic Res* 17:943–954
45. Ludy C, Lang A (2006) A 3-year field-scale monitoring of foliage-dwelling spiders (Araneae) in transgenic Bt maize fields and adjacent field margins. *Biol Control* 38:314–324
46. Ludy C, Lang A (2006) Bt maize pollen exposure and impact on the garden spider, *Araneus diadematus*. *Entomol Exp Appl* 118:145–156
47. Lundgren JG, Wiedenmann RN (2005) Tritrophic interactions among Bt (CryM1) corn, aphid prey, and the predator *Coleomegilla maculata* (Coleoptera: Coccinellidae). *Environ Entomol* 34:1621–1625
48. Dively GP (2005) Impact of transgenic VIP3a x Cry1ab lepidopteran-resistant field corn on the nontarget arthropod community. *Environ Entomol* 34:1267–1291
49. Guo JY, Wan FH, Dong L, Lovei GL, Han ZJ (2008) Tri-trophic interactions between Bt cotton, the herbivore *Aphis gossypii* Glover (Homoptera: Aphididae), and the predator *Chrysopa pallens* (Rambur) (Neuroptera: Chrysopidae). *Environ Entomol* 37:263–270
50. Naranjo SE (2005) Long-term assessment of the effects of transgenic Bt cotton on the abundance of nontarget arthropod natural enemies. *Environ Entomol* 34: 1193–1210
51. Naranjo SE (2005) Long-term assessment of the effects of transgenic Bt cotton on the function of the natural enemy community. *Environ Entomol* 34:1211–1223
52. Hagerty AM, Kilpatrick AL, Turnipseed SG, Sullivan MJ, Bridges WC (2005) Predaceous arthropods and lepidopteran pests on conventional, Bollgard, and Bollgard II cotton under untreated and disrupted conditions. *Environ Entomol* 34:105–114
53. Head G, Moar M, Eubanks M, Freeman B, Ruberson J, Hagerty A, Turnipseed S (2005) A multiyear, large-scale comparison of arthropod populations on commercially managed Bt and non-Bt cotton fields. *Environ Entomol* 34:1257–1266
54. Torres JB, Ruberson JR (2006) Interactions of Bt-cotton and the omnivorous big-eyed bug *Geocoris punctipes* (Say), a key predator in cotton fields. *Biol Control* 39:47–57
55. Torres JB, Ruberson JR (2008) Interactions of *Bacillus thuringiensis* Cry1Ac toxin in genetically engineered cotton with predatory heteropterans. *Transgenic Res* 17:345–354
56. Sisterson M, Biggs RW, Manhardt NM, Carriere Y, Dennehy TJ, Tabashnik BE (2007) Effects of transgenic Bt cotton on insecticide use and abundance of two generalist predators. *Entomol Exp Appl* 124:305–311
57. Reed GL, Jensen AS, Riebe J, Head G, Duan JJ (2001) Transgenic Bt potato and conventional insecticides for Colorado potato beetle management: comparative efficacy and non-target impacts. *Entomol Exp Appl* 100: 89–100
58. Duan JJ, Head G, Jensen A, Reed G (2004) Effects of transgenic *Bacillus thuringiensis* potato and conventional insecticides for Colorado potato beetle (Coleoptera: Chrysomelidae) management on the abundance of ground-dwelling arthropods in Oregon potato ecosystems. *Environ Entomol* 33:275–281
59. Dogan EB, Berry RE, Reed GL, Rossignol PA (1996) Biological parameters of convergent lady beetle (Coleoptera: Coccinellidae) feeding on aphids (Homoptera: Aphididae) on transgenic potato. *J Econ Entomol* 89:1105–1108
60. Armer CA, Berry RE, Kogan M (2000) Longevity of phytophagous heteropteran predators feeding on transgenic Bt-potato plants. *Entomol Exp Appl* 95:329–333
61. Ferry N, Mulligan EA, Majerus MEN, Gatehouse AMR (2007) Bitrophic and tritrophic effects of Bt Cry3A transgenic potato on beneficial, non-target, beetles. *Transgenic Res* 16:795–812
62. Bell HA, Down RE, Fitches EC, Edwards JP, Gatehouse AMR (2003) Impact of genetically modified potato expressing plant-derived insect resistance genes on the predatory bug *Podisus maculiventris*. *Biocontrol Sci Technol* 13:729–742
63. Graham J, Gordon SC, Smith K, McNicol RJ, McNicol JW (2002) The effect of the cowpea trypsin inhibitor in strawberry on damage by vine weevil under field conditions. *J Hort Sci Biotechnol* 77:33–40
64. Alvarez-Alfageme F, Martinez M, Pascual-Ruiz S, Castanera P, Diaz I, Ortego F (2007) Effects of potato plants expressing a barley cystatin on the predatory bug *Podisus maculiventris* via herbivorous prey feeding on the plant. *Transgenic Res* 16:1–13
65. Ferry N, Jouanin L, Ceci LR, Mulligan EA, Emami K, Gatehouse JA, Gatehouse AMR (2005) Impact of oilseed rape expressing the insecticidal serine protease inhibitor, mustard trypsin inhibitor-2 on the beneficial predator *Pterostichus madidus*. *Mol Ecol* 14:337–349
66. Jørgensen HB, Lövei GL (1999) Tri-trophic effect on predator feeding: consumption by the carabid *Harpalus affinis* of *Heliothis armigera* caterpillars fed on proteinase inhibitor-containing diet. *Entomol Exp Appl* 93:113–116

67. Burgess EPJ, Malone LA, Christeller JT, Lester MT, Murray C, Philip BA, Phung MM, Tregidga EL (2002) Avidin expressed in transgenic tobacco leaves confers resistance to two noctuid pests, *Helicoverpa armigera* and *Spodoptera litura*. *Transgenic Res* 11:185–198
68. Mulligan EA, Ferry N, Jouanin L, Romeis J, Gatehouse AMR (2010) Characterisation of adult green lacewing (*Chrysoperla carnea*) digestive physiology: comparison of the impact of genetically modified crops and conventional pest control. *Pest Manag Sci* 66(3):325–336
69. Mulligan EA, Ferry N, Jouanin L, Walters K, Port GR, Gatehouse AMR (2006) Comparing the impact of conventional pesticide and use of a transgenic pest-resistant crop on the beneficial carabid beetle *Pterostichus melanarius*. *Pest Manag Sci* 62:999–1012
70. Ferry N, Raemaekers RJM, Majerus MEN, Jouanin L, Port G, Gatehouse JA, Gatehouse AMR (2003) Impact of oilseed rape expressing the insecticidal cysteine protease inhibitor oryzacystatin on the beneficial predator *Harmonia axyridis* (Multicolored Asian Ladybeetle). *Mol Ecol* 12:493–504
71. Bouchard E, Cloutier C, Michaud D (2003) Oryzacystatin I expressed in transgenic potato induces digestive compensation in an insect natural predator via its herbivorous prey feeding on the plant. *Mol Ecol* 12:2439–2446
72. Lawo NC, Romeis J (2008) Assessing the utilization of a carbohydrate food source and the impact of insecticidal proteins on larvae of the green lacewing, *Chrysoperla carnea*. *Biol Control* 44:389–398
73. Prutz G, Dettner K (2004) Effect of Bt corn leaf suspension on food consumption by *Chilo partellus* and life history parameters of its parasitoid *Cotesia flavipes* under laboratory conditions. *Entomol Exp Appl* 111:179–187
74. Pilcher CD, Rice ME, Obrycki JJ (2005) Impact of transgenic *Bacillus thuringiensis* corn and crop phenology on five nontarget arthropods. *Environ Entomol* 34:1302–1316
75. Sanders CJ, Pell JK, Poppy GM, Raybould A, Garcia-Alonso M, Schuler TH (2007) Host-plant mediated effects of transgenic maize on the insect parasitoid *Campoletis sonorensis* (Hymenoptera: Ichneumonidae). *Biol Control* 40:362–369
76. Bernal JS, Griset JG, Gillogly PO (2002) Impacts of developing on Bt maize-intoxicated hosts on fitness parameters of a stem borer parasitoid. *J Entomol Sci* 37:27–40
77. Liu XX, Zhang QW, Xu BL, Li JC (2006) Effects of Cry1Ac toxin of *Bacillus thuringiensis* and nuclear polyhedrosis virus of *Helicoverpa armigera* (Hübner) (Lepidoptera: Noctuidae) on larval mortality and pupation. *Pest Manag Sci* 62:729–737
78. Liu XX, Zhang QW, Zhao JZ, Cai QN, Xu HL, Li JC (2005) Effects of the Cry1Ac toxin of *Bacillus thuringiensis* on *Microplitis mediator*, a parasitoid of the cotton bollworm, *Helicoverpa armigera*. *Entomol Exp Appl* 114:205–213
79. Baur ME, Boethel DJ (2003) Effect of Bt-cotton expressing Cry1A(c) on the survival and fecundity of two hymenopteran parasitoids (Braconidae, Encyrtidae) in the laboratory. *Biol Control* 26:325–332
80. Walker GP, Cameron PJ, MacDonald FM, Madhusudhan VV, Wallace AR (2007) Impacts of *Bacillus thuringiensis* toxins on parasitoids (Hymenoptera: Braconidae) of *Spodoptera litura* and *Helicoverpa armigera* (Lepidoptera: Noctuidae). *Biol Control* 40:142–151
81. Schuler TH, Denholm I, Clark SJ, Stewart CN, Poppy GM (2004) Effects of Bt plants on the development and survival of the parasitoid *Cotesia plutellae* (Hymenoptera: Braconidae) in susceptible and Bt-resistant larvae of the diamondback moth, *Plutella xylostella* (Lepidoptera: Plutellidae). *J Insect Physiol* 50:435–443
82. Chen M, Zhao JZ, Collins HL, Earle ED, Cao J, Shelton AM (2008) A critical assessment of the effects of Bt transgenic plants on parasitoids. *PLoS ONE* 3:1–7
83. Barraclough EI, Burgess EPJ, Philip BA, Wohlers MW, Malone LA (2009) Tritrophic impacts of Bt-expressing transgenic pine on the parasitoid *Meteorus pulchricornis* (Hymenoptera: Braconidae) via its host *Pseudocoremia suavis* (Lepidoptera: Geometridae). *Biol Control* 49:192–199
84. Marchetti E, Alberghini S, Battisti A, Squartini A, Baronio P, Dindo ML (2009) Effects of conventional and transgenic *Bacillus thuringiensis galleriae* toxin on *Exorista larvarum* (Diptera: Tachinidae), a parasitoid of forest defoliating Lepidoptera. *Biol Control Sci Technol* 19:463–473
85. Johnson MT, Gould F (1992) Interaction of genetically engineered host plant resistance and natural enemies of *Heliothis virescens* (Lepidoptera: Noctuidae) in tobacco. *Environ Entomol* 21:586–597
86. Bell HA, Fitches EC, Down RE, Ford L, Marris GC, Edwards JP, Gatehouse JA, Gatehouse AMR (2001) Effect of dietary cowpea trypsin inhibitor (CpTI) on the growth and development of the tomato moth *Lacanobia oleracea* (Lepidoptera: Noctuidae) and on the success of the gregarious ectoparasitoid *Eulophus pennicornis* (Hymenoptera: Eulophidae). *Pest Manag Sci* 57:57–65
87. Bell HA, Kirkbride-Smith AE, Marris GC, Edwards JP, Gatehouse AMR (2004) Oral toxicity and impact on fecundity of three insecticidal proteins on the gregarious ectoparasitoid *Eulophus pennicornis* (Hymenoptera: Eulophidae). *Agric Forest Entomol* 6:215–222
88. Geng JH, Shen ZR, Song K, Zheng L (2006) Effect of pollen of regular cotton and transgenic Bt plus CpTI cotton on the survival and reproduction of the parasitoid wasp *Trichogramma chilonis* (Hymenoptera: Trichogrammatidae) in the laboratory. *Environ Entomol* 35:1661–1668
89. Azzouz H, Cherqui A, Campan EDM, Rahbe Y, Duport G, Jouanin L, Kaiser L, Giordanengo P (2005) Effects of plant protease inhibitors, oryzacystatin I and soybean Bowman-Birk inhibitor, on the aphid *Macrosiphum euphorbiae* (Homoptera, Aphididae) and its parasitoid *Aphelinus abdominalis* (Hymenoptera, Aphelinidae). *J Insect Physiol* 51:75–86

90. Ashouri A, Michaud D, Cloutier C (2001) Recombinant and classically selected factors of potato plant resistance to the Colorado potato beetle, *Leptinotarsa decemlineata*, variously affect the potato aphid parasitoid *Aphidius nigripes*. *Biocontrol* 46:401–418
91. Schuler TH, Denholm I, Jouanin L, Clark SJ, Clark AJ, Poppy GM (2001) Population-scale laboratory studies of the effect of transgenic plants on nontarget insects. *Mol Ecol* 10:1845–1853
92. Cowgill SE, Danks C, Atkinson HJ (2004) Multitrophic interactions involving genetically modified potatoes, nontarget aphids, natural enemies and hyperparasitoids. *Mol Ecol* 13:639–647
93. Wakefield ME, Bell HA, Gatehouse AMR (2010) Longevity and fecundity of *Eulophus pennicornis*, an ectoparasitoid of the tomato moth *Lacanobia oleracea*, is affected by nutritional state and diet quality. *Agric Forest Entomol* 12:19–27
94. Couty A, de la Vina G, Clark SJ, Kaiser L, Pham-Delegue MH, Poppy GM (2001) Direct and indirect sublethal effects of *Galanthus nivalis* agglutinin (GNA) on the development of a potato-aphid parasitoid, *Aphelinus abdominalis* (Hymenoptera: Aphelinidae). *J Insect Physiol* 47:553–561
95. Couty A, Poppy GM (2001) Does host-feeding on GNA-intoxicated aphids by *Aphelinus abdominalis* affect their longevity and/or fecundity? *Entomol Exp Appl* 100: 331–337
96. Romeis J, Babendreier D, Wackers FL (2003) Consumption of snowdrop lectin (*Galanthus nivalis* agglutinin) causes direct effects on adult parasitic wasps. *Oecologia* 134: 528–536
97. Hogervorst PAM, Wackers FL, Woodring J, Romeis J (2009) Snowdrop lectin (*Galanthus nivalis* agglutinin) in aphid honeydew negatively affects survival of a honeydew-consuming parasitoid. *Agric Forest Entomol* 11:161–173
98. Setamou M, Bernal JS, Legaspi JC, Mirkov TE (2002) Effects of snowdrop lectin (*Galanthus nivalis* agglutinin) expressed in transgenic sugarcane on fitness of *Cotesia flavipes* (Hymenoptera: Braconidae), a parasitoid of the nontarget pest *Diatraea saccharalis* (Lepidoptera: Crambidae). *Ann Entomol Soc Am* 95:75–83
99. Setamou M, Bernal JS, Legaspi JC, Mirkov TE, Legaspi BC (2002) Evaluation of lectin-expressing transgenic sugarcane against stalkborers (Lepidoptera: Pyralidae): Effects on life history parameters. *J Econ Entomol* 95:469–477
100. Bell HA, Fitches EC, Down RE, Marris GC, Edwards JP, Gatehouse JA, Gatehouse AMR (1999) The effect of snowdrop lectin (GNA) delivered via artificial diet and transgenic plants on *Eulophus pennicornis* (Hymenoptera: Eulophidae), a parasitoid of the tomato moth *Lacanobia oleracea* (Lepidoptera: Noctuidae). *J Insect Physiol* 45:983–991
101. Tomov BW, Bernal JS, Vinson SB (2003) Impacts of transgenic sugarcane expressing GNA lectin on parasitism of Mexican rice borer by *Parallorhogas pyralophagus* (Marsh) (Hymenoptera: Braconidae). *Environ Entomol* 32:866–872
102. Wakefield ME, Bell HA, Fitches EC, Edwards JP, Gatehouse AMR (2006) Effects of *Galanthus nivalis* agglutinin (GNA) expressed in tomato leaves on larvae of the tomato moth *Lacanobia oleracea* (Lepidoptera: Noctuidae) and the effect of GNA on the development of the endoparasitoid *Meteorus gyrator* (Hymenoptera: Braconidae). *Bull Entomol Res* 96:43–52
103. Romeis J, Hellmich RL, Candolfi MP, Carstens K, De Schrijver A, Gatehouse AMR, Herman RA, Huesing, JE, McLean MA, Raybould A, Shelton AM, Waggoner A (2010) Recommendations for the design of laboratory studies on non-target arthropods for risk assessment of genetically engineered plants. *Transgenic Res* 1–22
104. Marvier M, McCreedy C, Regetz J, Kareiva P (2007) A meta-analysis of effects of Bt cotton and maize on nontarget invertebrates. *Science* 316:1475–1477
105. Wolfenbarger LL, Naranjo SE, Lundgren JG, Bitzer RJ, Watrud LS (2008) Bt crop effects on functional guilds of non-target arthropods: a meta-analysis. *PLoS ONE* 3(5):e2118. doi:10.1371/journal.pone.0002118
106. Duan JJ, Lundgren JG, Naranjo S, Marvier M (2010) Extrapolating non-target risk of Bt crops from laboratory to field. *Biol Lett* 23(6):74–77
107. Sanvido O, Romeis J, Bigler F (2007) Ecological impacts of genetically modified crops: Ten years of field research and commercial cultivation. In: *Green Gene Technology: Research in an Area of Social Conflict*, vol 107. *Advances in Biochemical Engineering/Biotechnology*. Springer-Verlag Berlin, Berlin, pp 235–278
108. Malone MA, Burgess EPJ (2009) Impact of genetically modified crops on pollinators. In: Ferry N, Gatehouse AMR (eds) *Environmental impact of genetically modified crops*. CAB International, Wallingford, pp 199–224
109. Young BG (2006) Changes in herbicide use patterns and production practices resulting from glyphosate-resistant crops. *Weed Technol* 20:301–307
110. Service RF (2007) Glyphosate – the conservationist's friend? *Science* 316:1116–1117
111. Foresman C, Glasgow L (2008) US grower perceptions and experiences with glyphosate-resistant weeds. *Pest Manag Sci* 64:388–391
112. Sankula S (2006) Quantification of the impacts on US agriculture of biotechnology-derived crops planted in 2005. *Executive Summary*. National Center for Food and Agricultural Policy, Washington, DC
113. Dill GM, CaJacob CA, Padgett SR (2008) Glyphosate-resistant crops: adoption, use and future considerations. *Pest Manag Sci* 64:326–331
114. Gianessi LP, Reigner NP (2007) The value of herbicides in U.S. crop production. *Weed Technol* 21:559–566
115. Owen MDK (2005) Maize and soybeans – controllable volunteerism without ferality. In: Gressel J (ed) *Crop ferality and volunteerism*. CRC, Boca Raton, pp 149–165
116. Johnson KL, Raybould AF, Hudson MD, Poppy GM (2006) How does scientific risk assessment of GM crops fit within the wider risk analysis? *Trends Plant Sci* 12:1–5

117. Madsen KH, Sandoe P (2005) Ethical reflections on herbicide-resistant crops. *Pest Manag Sci* 61:318–325
118. Dill GM (2005) Glyphosate-resistant crops; history, status and future. *Pest Manag Sci* 61:219–224
119. Gianessi LP (2005) Economic and herbicide use impacts of glyphosate-resistant crops. *Pest Manag Sci* 61:241–245
120. Rissler J, Mellon M (1996) The ecological risks of engineered organisms. MIT, Cambridge
121. Altieri MA (2000) The ecological impacts of transgenic crops on agroecosystem health. *Agroecology in action*. University of California: Berkeley
122. Vidal RA, Trezzi MM, De Prado R, Ruiz-Santaella JP, Vial-Aiub M (2007) Glyphosate resistant biotypes of wild poinsettia (*Euphorbia heterophylla* L.) and its risk analysis on glyphosate-resistant soybeans. *J Food Agric Environ* 5:265–269
123. Werth JA, Preston C, Taylor IN, Charles GW, Roberts GN, Baker J (2008) Managing the risk of glyphosate resistance in Australian glyphosate-resistant cotton production systems. *Pest Manag Sci* 64:417–421
124. Krebs JR, Wilson JD, Bradbury RB, Siriwardena GM (1999) The second Silent Spring? *Nature* 400:611–612
125. Schutte G (2003) Herbicide resistance: Promises and prospects of biodiversity for European agriculture. *Agric Hum Values* 20:217–230
126. Benton TG (2007) Managing farming's footprint on biodiversity. *Science* 315:341–342
127. Davis LC (2006) Genetic engineering, ecosystem change, and agriculture: an update. *Biotechnol Mol Biol Rev* 1:87–102
128. Norris RF (2005) Ecological bases of interactions between weeds and organisms in other pest categories. *Weed Sci* 53:909–913
129. Scursoni J, Peterson D, Forcella F, Gunsolus J, Arnold RB, Owen M, Smeda OR, Vidrine R (2001) Weed diversity in glyphosate-tolerant soybean from Minnesota to Louisiana. In: Hartzler RG (ed) CD-ROM Computer File. North Central Weed Science Society, Champaign, Milwaukee
130. Bitzer RJ, Buckelew LD, Pedigo LP (2002) Effects of transgenic herbicide-resistant soybean varieties and systems on surface-active springtails (Entognatha: Collembola). *Environ Entomol* 31:449–461
131. Freckleton RP, Stephens PA, Sutherland WJ, Watkinson AR (2004) Amelioration of biodiversity impacts of genetically modified crops: predicting transient versus long-term effects. *Proc R Soc Biol* 271:325–331
132. Cerdeira AL, Duke SO (2006) The current status and environmental impacts of glyphosate-resistant crops: a review. *J Environ Qual* 35:1633–1658
133. Scursoni JA, Forcella F, Gunsolus J (2007) Weed escapes and delayed emergence in glyphosate-resistant soybean. *Crop Protect* 26:212–218
134. Heard MS, Hawes C, Champion GT, Clark SJ, Firbank LG, Haughton AJ, Parish AM, Perry JN, Rothery JN, Scott P, Skellern RJ, Squire MP, Hill GR (2003) Weeds in fields with contrasting conventional and genetically modified herbicide-tolerant crops. I. Effects on abundance and diversity. *Philos Trans R Soc Lond B* 358:1819–1832
135. Heard MS, Clark SJ, Rothery P, Perry JN, Bohan DA, Brooks DR, Champion GT, Dewar AM, Hawes C, Haughton AJ, May MJ, Scott RJ, Stuart RS, Squire GR, Firbank LG (2006) Effects of successive seasons of genetically modified herbicide-tolerant maize cropping on weeds and invertebrates. *Ann Appl Biol* 149:249–254
136. Roy DB, Bohan DA, Haughton AJ, Hill MO, Osborne JL, Clark SJ, Perry JN, Rothery P, Scott RJ, Brooks DR, Champion GT, Hawes C, Heard MS, Firbank LG (2003) Invertebrates and vegetation of field margins adjacent to crops subject to contrasting herbicide regimes in the Farm Scale Evaluations of genetically modified herbicide-tolerant crops. *Philos Trans R Soc Lond B* 358:1879–1898
137. Owen MDK (2000) Current use of transgenic herbicide-resistant soybean and corn in the USA. *Crop Prot* 19: 765–771
138. Luan V, Figueroa SJ, Baltazar MB, Gomez MR, Townsend LR, Schoper JB (2001) Maize pollen longevity and distance isolation requirements for effective pollen control. *Crop Sci* 41:1551–1557
139. Palmer RG, Gai J, Sun H, Burton JW (2001) Production and evaluation of hybrid soybean. In: Janick J (ed) *Plant breeding reviews*. Wiley, New York, pp 263–308
140. Westgate ME, Lizaso J, Batchelor W (2003) Quantitative relationship between pollen shed density and grain yield in maize. *Crop Sci* 43:934–942
141. Abud SPI, Moreira MS, Farias Neto CT, Vianna AL, Andrade GR, Nunes SRM, Guerezoni J, Monteiro RA, Assuncao PMFO, Rech MS, Aragao EL (2004) Distance of flow in transgenic BROO-69515 and the non transgenic soybean in the Cerrado Region, Brasil. In: Moscardi F, Hoffmann-Campo CB, Saraiva OF, Galerani PR, Krzyzanowski FC, Carrao-Panizzi MC (eds) VII World Soybean Research Conference. Brazilian Agricultural Research Corporation, Foz do Iguassu, Brazil, pp 110–111
142. Matus-Cadiz PH, Horak MJ, Blomquist LK (2004) Gene flow in wheat at the field scale. *Crop Sci* 44:718–727
143. Charles D (2007) U.S. courts say transgenic crops need tighter scrutiny. *Science* 315:1069
144. Fisher L (2007) Growers continue to grow and use Roundup Ready alfalfa but Monsanto Company is disappointed with preliminary injunction affecting purchase and planting. <http://www.monsanto.com> Accessed 8 March 2007
145. Harriman P (2007) Roundup Ready concerns land alfalfa seed in court. http://4.26.159.139/static/news/NEWSID_8242.php. Accessed 8 March 2007
146. Weise E (2007) Effect of genetically engineered alfalfa cultivate a debate, USA Today. http://www.columbia.org/pdf_files/centerforfoodsafety20.pdf. Accessed 08 December 2010
147. Hurburgh JCR (2000) The GMO controversy and grain handling for 2000. GMO and Grain, Iowa State University, Ames
148. Hurburgh JCR (2003) Constraints for isolation and traceability of grains. Iowa State University, Ames

149. Ginder RG (2001) Channelling, identity preservation and the value chain: Lessons from the recent problems with starlink corn, Staff General Research Papers, Iowa State University, Department of Economics, <http://econpapers.repec.org/RePEc:isu:genres:11480>
150. Qaim M (2010) The benefits of genetically modified crops—and the costs of inefficient regulation. Resources for the future. www.rff.org/Publications/WPC/Pages/The-Benefits-of-Genetically-Modified-Crops-and-the-Costs-of-Inefficient-Regulation.aspx
151. Losey JE, Rayer LS, Carter ME (1999) Transgenic pollen harms monarch larvae. *Nature* 399:214
152. Stanley-Horn DE, Dively GP, Hellmich RL, Mattila HR, Sears MK, Rose R, Jesse LCH, Losey JE, Obrycki JJ, Lewis L (2001) Assessing the impact of Cry1Ab-expressing corn pollen on monarch butterfly larvae in field studies. *Proc Natl Acad Sci USA* 98:11931–11936
153. Zangerl AR, McKenna D, Wraight CL, Carroll M, Ficarelllo P, Warner R, Berenbaum MR (2001) Effects of exposure to event 176 *Bacillus thuringiensis* corn pollen on monarch and black swallowtail caterpillars under field conditions. *Proc Natl Acad Sci USA* 98:11908–11912
154. Pleasants JM, Hellmich RL, Dively GP, Sears MK, Stanley-Horn DE, Mattila HR, Foster JE, Clark P, Jones GD (2001) Corn pollen deposition on milkweeds in and near cornfields. *Proc Natl Acad Sci USA* 98:11919–11924
155. Oberhauser KS, Prysby MD, Mattila HR, Stanley-Horn DE, Sears MK, Dively G, Olson E, Pleasants JM, Lam WKF, Hellmich RL (2001) Temporal and spatial overlap between monarch larvae and corn pollen. *Proc Natl Acad Sci USA* 98:11913–11918
156. Sears MK, Hellmich RL, Stanley-Horn DE, Oberhauser KS, Pleasants JM, Mattila HR, Siegfried BD, Dively GP (2001) Impact of *Bt* corn pollen on monarch butterfly populations: a risk assessment. *Proc Natl Acad Sci USA* 98:11937–11942
157. Gatehouse AMR, Ferry N, Raemaekers RJM (2002) The case of the monarch butterfly: a verdict is returned. *Trends Genet* 18:249–251
158. Obrist LB, Dutton A, Romeis J, Bigler F (2006) Biological activity of Cry1Ab toxin expressed by Bt maize following ingestion by herbivorous arthropods and exposure of the predator *Chrysoperla carnea*. *Biocontrol* 51:31–48
159. Ellstrand NC (2003) Current knowledge of gene flow in plants: implications for transgene flow. *Philos Trans R Soc Lond B* 358:1163–1170
160. Hall LM, Good A, Beckie HJ, Warwick SI (2003) Gene flow in herbicide-resistant canola (*Brassica napus*): the Canadian experience. In: Lelley T, Balazs E, Tepfer M (eds) *Ecological impact of GMO dissemination in agroecosystems*. Facultas Verlagsun Buchhandels AG, Austria, pp 57–66
161. Clark EA (2006) Environmental risks of genetic engineering. *Euphytica* 148:47–60
162. Mallory-Smith C, Zapiola M (2008) Gene flow from glyphosate-resistant crops. *Pest Manag Sci* 64:428–440

Transgenic Crops, Next Generation

MARTINA NEWELL-MCGLOUGHLIN

Department of Plant Pathology, University of California Davis, Davis, CA, USA

Article Outline

Glossary

Definition of the Subject

Introduction

Progress to Date

The Next Generation

Barriers to Introduction

Future Directions

Bibliography

Glossary

Abiotic stress External (nonliving) factors which can cause harmful effects to plants, such as soil conditions, drought, and extreme temperatures.

Acclimatization Adaptation of an organism to a new environment.

Adaptation In the evolutionary sense, some heritable feature of an individual's phenotype that improves its chances of survival and reproduction in the existing environment.

Additive genetic variance Genetic variance associated with the average effects of substituting one allele for another.

Agronomic performance/trait Pertains to practices of agricultural production and its costs and the management of crop land. Examples of agronomic traits include yield, input requirements, stress tolerance.

Agrobacterium tumefaciens A bacterium normally responsible for causing crown gall disease in a variety of plants. A plasmid has been isolated from this bacterium that is useful in plant genetic engineering. This plasmid, called the Ti plasmid, has been modified so that it does not cause disease but can carry foreign DNA into susceptible plant cells.

Aldolase An enzyme, not subject to allosteric regulation, that catalyzes in a reversible reaction the

cleavage of fructose 1,6-biphosphate to form dihydroxyacetone phosphate and glyceraldehyde 3-phosphate. The enzyme catalyzing the fourth reaction in the glycolytic pathway, which splits a monosaccharide into two 3-carbon units.

Allele Any of several alternative forms of a given gene.

Allele frequency Often called gene frequency. A measure of how common an allele is in a population the proportion of all alleles at one gene locus that are of one specific type in a population.

Allelic exclusion A process whereby only one immunoglobulin light chain and one heavy chain gene are transcribed in any one cell; the other genes are repressed.

Allogenic Of the same species, but with a different genotype.

Allopolyploid Polyploid produced by the hybridization of two species.

Allopolyploid Plants Plants having more than two sets of haploid chromosomes inherited from different species.

Allosteric Regulation Regulation of an enzyme's activity by binding of a small molecule at a site that does not overlap the active site region.

Anabolic The part of metabolism that is concerned with synthetic reactions.

Aneuploid Having a chromosome number that is not an exact multiple of the haploid number, caused by one chromosome set being incomplete or chromosomes being present in extra numbers.

Aneuploidy The condition of a cell or an organism that has additions or deletions of a small number of whole chromosomes from the expected balanced diploid number of chromosomes.

Antibiotic Chemical substance formed as a metabolic by-product in bacteria or fungi and used to treat bacterial infections. Antibiotics can be produced naturally, using microorganisms, or synthetically.

Antibody A protein produced by the immune system in response to an antigen (a molecule that is perceived to be foreign). Antibodies bind specifically to their target antigen to help the immune system destroy the foreign entity.

Antinutrients Substances that act in direct competition with or otherwise inhibit or interfere with the use or absorption of a nutrient.

Antisense RNA RNA produced by copying and reversing a portion of an RNA-encoding DNA, usually including a protein-specifying region, and placing it next to a transcription-control sequence. This cassette can be delivered to the target cell, resulting in genetic transformation and production of RNA that is complementary to the RNA that is produced from the original, not reversed, DNA segment. This complementary, or antisense, RNA is able to bind to the complementary sequences of the target RNA, resulting in the inhibition of expression of the target gene.

Antiserum Blood serum containing specific antibodies against an antigen. Antisera are used to confer passive immunity to diseases and as analytical and preparative reagents for antigens, for example, to determine potential allergenicity.

Avirulent Unable to cause disease.

Bacillus thuringiensis (Bt) A naturally occurring microorganism which produces a toxin protein that only kills organisms with alkaline stomachs, such as insect larvae. As and when delivered as a part of the whole killed organism, this toxin protein has been used for biological control for decades. The genetic information that encodes the toxin protein was identified and was moved into plants to make them insect tolerant.

Bioconversion Chemical restructuring of raw materials by using a biocatalyst.

Biodegradable Capable of being broken down by the action of microorganisms, usually by microorganisms and under conditions generally in the environment.

Bioinformatics The discipline encompassing the development and utilization of computational facilities to store, analyze, and interpret biological data.

Biomass The totality of biological matter in a given area. As commonly used in biotechnology, it refers to the use of cellulose, a renewable resource, for the production of chemicals that can be used to generate energy or as alternative feedstocks for the chemical industry to reduce dependence on nonrenewable fossil fuels.

Bioprocess A process in which living cells, or components thereof, are used to produce a desired end product.

- Biosynthesis** Production of a chemical by a living organism.
- Biotechnology** Development of products by a biological process. Production may be carried out by using intact organisms, such as yeasts and bacteria, or by using natural substances (e.g., enzymes) from organisms.
- Biosynthetic** The formation of complex compounds from simple substances by living organisms.
- Biotic Stress** Living organisms which can harm plants, such as viruses, fungi, and bacteria, and harmful insects. (See Abiotic stress).
- Callus** A cluster of undifferentiated plant cells that can, for some species, be induced to form the whole plant.
- Calvin Cycle** A series of enzymatic reactions, occurring during photosynthesis, in which glucose is synthesized from carbon dioxide.
- Catalyst** An agent (such as an enzyme or a metallic complex) that facilitates a reaction but is not itself changed at the completion of the reaction.
- Catabolic** The part of metabolism that is concerned with degradation reactions.
- Chloroplast** A chlorophyll-containing photosynthetic organelle, found in eukaryotic cells that can harness light energy.
- Cistron** A length of chromosomal DNA representing the smallest functional unit of heredity, essentially identical to a gene.
- Clone** A group of genes, cells, or organisms derived from a common ancestor. Because there is no combining of genetic material (as in sexual reproduction), the members of the clone are genetically identical or nearly identical to the parent.
- Codon** A sequence of three nucleotide bases that in the process of protein synthesis specifies an amino acid or provides a signal to stop or start protein synthesis (translation).
- Coenzyme** An organic compound that is necessary for the functioning of an enzyme. Coenzymes are smaller than the enzymes themselves and may be tightly or loosely attached to the enzyme protein molecule.
- Cofactor** A nonprotein substance required for certain enzymes to function. Cofactors can be coenzymes or metallic ions.
- Comparative genomics** The comparison of genome structure and function across different species for further understanding of biological mechanisms and evolutionary processes.
- Composition analysis** The determination of the concentration of compounds in a plant. Compounds that are commonly quantified are proteins, fats, carbohydrates, minerals, vitamins, amino acids, fatty acids, and antinutrients.
- Conventional breeding** Breeding of plants carried out by controlled transfer of pollen from one plant to another followed by selection of progeny through multiple generations for a desirable phenotype. This method has also often included irradiation or mutation of plants or seeds to induce extra variation in the donor material.
- Complementary DNA (cDNA)** DNA synthesized from an expressed messenger RNA through a process known as reverse transcription. This type of DNA is used for cloning or as a DNA probe for locating specific genes in DNA hybridization studies.
- Coumarins** White vanilla-scented crystalline esters used in perfumes and flavorings and as an anticoagulant. Formula: $C_9H_6O_2$.
- Crossbreeding** Interbreeding (of animals or plants) using parents of different races, varieties, breeds, etc.
- Cyto-** A prefix referring to cell or cell plasm.
- Cytokines** Intercellular signals, usually protein or glycoprotein, involved in the regulation of cellular proliferation and function.
- Diet** A specific allowance or selection of food or feed that a person or animal regularly consumes.
- Diploid** A cell with two complete sets of chromosomes. Cf. Haploid.
- DNA sequencing** Technologies through which the order of base pairs in a DNA molecule can be determined.
- Enzyme** A protein catalyst that facilitates specific chemical or metabolic reactions necessary for cell growth and reproduction. Cf. Catalyst.
- Epigenetics** The study of changes in gene expression caused by mechanisms other than changes in the underlying DNA sequence – hence the name *epi-* (Greek: *επί-* over, above, outer) -genetics. Examples of such changes might be DNA methylation or

histone deacetylation, both of which serve to suppress gene expression without altering the sequence of the silenced genes.

Event The term used to describe a plant and its offspring that contain a specific insertion of DNA. Such events will be distinguishable from other events by their unique site of integration of the introduced DNA.

Exposure assessment The qualitative and/or quantitative evaluation of the likely exposure to biological, chemical, and physical agents via different sources.

Feedstock The raw material used in chemical or biological processes.

Flavonoids Any of a group of organic compounds that occur as pigments in fruit and flowers.

Food additive Any substance not normally consumed as a food by itself and not normally used as a typical ingredient of food, whether or not it has nutritive value, the intentional addition of which to a food for a technological (including organoleptic) purpose in the manufacture, processing, preparation, treatment, packing, packaging, transport, or holding of such food results, or may be expected to result (directly or indirectly), in it or its by-products becoming a component of or otherwise affecting the characteristics of such foods. The term does not include “contaminants” or substances added to food for maintaining or improving nutritional qualities.

Fructan A type of polymer of fructose, present in certain fruits.

Functional foods The Institute of Medicine’s Food and Nutrition Board defined functional foods as “any food or food ingredient that may provide a health benefit beyond the traditional nutrients it contains.”

Functional genomics The development and implementation of technologies to characterize the mechanisms through which genes and their products function and interact with each other and with the environment. The science that is usually applied to the studies of gene and the expression (mRNA) of usually large numbers of genes simultaneously.

Gene expression The process through which a gene is activated at particular time and place so that its functional product is produced.

Gene flow The exchange of genetic traits between populations by movement of individuals, gametes, or spores. It involves the spread of new variants among different populations through dispersal.

Gene silencing (See RNAi) A method usually performed by the expression of an mRNA of complementary or the same nucleotide sequence in a cell such that the expression of the mRNA causes the downregulation of the protein which is being targeted.

Gene transfer The transfer of genes to an organism. Usually used in terms of transfer of a gene to an organism other than the original organism, through the tools of biotechnology.

Gene A segment of chromosome that encodes the necessary regulatory and sequence information to direct the synthesis of a protein or RNA product. (See also Operator Regulatory g. Structural g. Suppressor g).

Gene mapping Determination of the relative locations of genes on a chromosome.

Gene sequencing Determination of the sequence of nucleotide bases in a strand of DNA.

Genetic engineering A technology used to alter the genetic material of living cells in order to make them capable of producing new substances or performing new functions.

Genetic map A Map showing the positions of genetic markers along the length of a chromosome relative to each other (genetic map) or in absolute distances from each other (physical map).

Genome The total hereditary material of a cell, comprising the entire chromosomal set found in each nucleus of a given species.

Genomics Science that studies the genomes (i.e., the complete genetic information) of living beings. This commonly entails the analysis of DNA sequence data and the identification of genes.

Genotype Genetic makeup of an individual or group. Cf. Phenotype.

Germplasm The total genetic variability, represented by germ cells or seeds, available within a particular population of organisms.

Gene pool The total genetic information contained within a given population.

Glycoalkaloid toxins Steroid-like compounds produced by plant members of the botanical family Solanaceae, most notably “solanine” present in potato tubers.

Golden rice In 1999, Swiss and German scientists announced the development of a genetically engineered rice crop that produces beta-carotene, a substance which the body converts to vitamin A. This improved nutrient rice was developed to treat individuals suffering from vitamin A deficiency, a condition that afflicts millions of people in developing countries, especially children and pregnant women.

Haploid A cell with half the usual number of chromosomes, or only one chromosome set. Sex cells are haploid. Cf. Diploid.

Hazard characterization The qualitative and/or quantitative evaluation of the nature of the adverse health effects associated with biological, chemical, and physical agents. For chemical agents, a dose-response assessment should be performed if the data are obtainable.

Hazard identification The identification of biological, chemical, and physical agents capable of causing adverse health or environmental effects.

Hazard A biological, chemical, or physical agent, or condition, with the potential to cause an adverse health or environmental effect.

Hereditary Capable of being transferred as genetic information from parent cells to progeny.

Heterozygote With respect to a particular gene at a defined chromosomal locus, a heterozygote has a different allelic form of the gene on each of the two homologous chromosomes.

Homologous Corresponding or alike in structure, position, or origin.

Homologous recombination Rearrangement of related DNA sequences on a different molecule by crossing over in a region of identical sequence.

Homozygote With respect to a particular gene at a defined chromosomal locus, a homozygote has the same allelic form of the gene on each of the two homologous chromosomes.

Hormone A chemical that acts as a messenger or stimulatory signal, relaying instructions to stop or start certain physiological activities. Hormones are synthesized in one type of cell and then released to direct the function of other cell types.

Horizontal gene transfer Transmission of DNA between species, involving close contact between the donor's DNA and the recipient, uptake of DNA by the recipient, and stable incorporation of the DNA into the recipient's genome.

Host A cell or organism used for growth of a virus, plasmid, or other form of foreign DNA, or for the production of cloned substances.

Hybridization Production of offspring, or hybrids, from genetically dissimilar parents. The process can be used to produce hybrid plants (by cross-breeding two different varieties) or hybridomas (hybrid cells formed by fusing two unlike cells, used in producing monoclonal antibodies). The term is also used to refer to the binding of complementary strands of DNA or RNA.

Hybrid The offspring of two parents differing in at least one genetic characteristic (trait). Also, a heteroduplex DNA or DNA-RNA molecule.

Identity preservation The segregation of one crop type from another at every stage from production and processing to distribution. This process is usually performed through audits and site visits and provides independent third-party verification of the segregation.

Immunoassay Technique for identifying substances based on the use of antibodies.

Immunogen Any substance that can elicit an immune response, especially specific antibody production. An immunogen that reacts with the elicited antibody may be called an antigen.

Inbred Progeny produced as a result of inbreeding.

Inducer A molecule or substance that increases the rate of enzyme synthesis, usually by blocking the action of the corresponding repressor.

Inserted DNA The segment of DNA that is introduced into the chromosome, plasmid, or other vectors using recombinant DNA techniques.

Introgressed Backcrossing of hybrids of two plant populations to introduce new genes into a wild population.

Inulins A fructose polysaccharide present in the tubers and rhizomes of some plants. Formula: $(C_6H_{10}O_5)_n$.

In vitro Literally, “in glass.” Performed in a test tube or other laboratory apparatus.

In vivo In the living organism.

Invertase activity Enzyme activity occurring in the intestinal juice of animals and in yeasts, that hydrolyses sucrose to glucose and fructose.

Isoflavones Water-soluble chemicals, also known as phytoestrogens, found in many plants and so named because they cause effects in the mammalian body somewhat similar to those of estrogen. The most investigated natural isoflavones, genistein and daidzen, are found in soy products and the herb red clover.

Knock in Replacement of a gene by a mutant version of the same gene using homologous recombination.

Knock out Inactivation of a gene by homologous recombination following transfection with a suitable DNA construct.

Linkage The tendency for certain genes to be inherited together due to their physical proximity on the chromosome.

Locus (Plural loci) The position of a gene, DNA marker, or genetic marker on a chromosome. (See gene locus).

Macronutrient Any substance, such as carbon, hydrogen, or oxygen, that is required in large amounts for healthy growth and development.

Marker Any genetic element (locus, allele, DNA sequence, or chromosome feature) which can be readily detected by phenotype, cytological or molecular techniques, and used to follow a chromosome or chromosomal segment during genetic analysis.

Marker assisted selection or marker aided selection (MAS) A process whereby a marker (morphological, biochemical, or one based on DNA/RNA variation) is used for indirect selection of a genetic determinant or determinants of a trait of interest (i.e., productivity, disease resistance, abiotic stress tolerance, and/or quality). This process is used in plant and animal breeding.

Mass spectrometry Analytical technique by which compounds in a vacuum compartment are ionized, eventually fragmented, accelerated, and detected based upon the mass-dependent behavior of the ionized compounds or their fragments in response to the application of a magnetic or electric field in a vacuum.

Messenger RNA (mRNA) Nucleic acid that carries instructions to a ribosome for the synthesis of a particular protein.

Metabolism All biochemical activities carried out by an organism to maintain life.

Metabolite A substance produced during or taking part in metabolism.

Metabolomics “Open-ended” analytical techniques that generate profiles of the metabolites, that is, chemical substances within a biological sample. Commonly differences between profiles of different (groups of) samples are determined and the identity of the associated metabolites elucidated. Contrary to targeted analysis, these techniques are indiscriminate in that they do not require prior knowledge of every single substance that is present.

Microarray A microscopic, ordered array of nucleic acids, proteins, small molecules, cells, or other substances that enables parallel analysis of complex biochemical samples. There are many different types of microarrays both from a biological and production system perspective. The generic terms “DNA array,” “GeneChip™,” or “hybridization array” are used to refer broadly to all types of oligonucleotide-based arrays. The two most common are cDNA arrays and genomic arrays. cDNA array: A microarray composed of grid of nucleic acid molecules of known composition linked to a solid substrate, which can be probed with total messenger RNA from a cell or tissue to reveal changes in gene expression relative to a control sample.

Micronutrient Any substance, such as a vitamin or trace element, essential for healthy growth and development but required only in minute amounts.

Mini-chromosome Contains only centromeres and telomeres with little additional DNA. This provides the ability to accept multiple genes coding for stacked traits. They are particularly useful because they allow scientists to add numerous genes onto one mini-chromosome and manipulate those genes easily because they are all in one place.

mRNA Messenger RNA.

Multigenic Of hereditary characteristics one that is specified by several genes.

Mutant A cell that manifests new characteristics due to a change in its DNA.

Mutation A structural change in a DNA sequence resulting from uncorrected errors during DNA replication.

Mutation Breeding Genetic change caused by natural phenomena or by use of mutagens. Stable mutations in genes are passed on to offspring unstable mutations are not.

Nitrogen fixation A biological process (usually associated with plants) whereby certain bacteria convert nitrogen in the air to ammonia, thus forming a nutrient essential for growth.

Nucleic acid Large molecules, generally found in the cell nucleus and/or cytoplasm, that are made up of nucleotide bases. The two kinds of nucleic acid are DNA and RNA.

Nucleotides The building blocks of nucleic acids. Each nucleotide is composed of sugar, phosphate, and one of four nitrogen bases. If the sugar is ribose, the nucleotide is termed a “ribonucleotide,” whereas deoxyribonucleotides have deoxyribose as the sugar component (i.e., adenine, cytosine, guanine, and thymine in the case of DNA). The sequence of the nucleotides within the nucleic acid determines, for example, the amino acid sequence of an encoded protein.

Nucleus In eukaryotic cells, the centrally located organelle that encloses most of the chromosomes. Minor amounts of chromosomal substance DNA are found in some other organelles, most notably the mitochondria and the chloroplasts.

Nutritionally improved Improving the quantity, ratio, and/or bioavailability of essential macro- and micronutrients and other compounds for which the clinical and epidemiological evidence is clear that they play a significant role in the maintenance of optimal health and are limiting in diets.

Nutraceutical The term was coined by the Foundation for Innovation in Medicine in 1991 and is defined as “any substance that may be considered a food or part of a food and provides medical or health benefits, including the prevention and treatment of disease.”

Organoleptic Able to perceive a sensory stimulus such as taste.

Operon Sequence of genes responsible for synthesizing the enzymes needed for biosynthesis of a molecule. An operon is controlled by an operator gene and a repressor gene.

Pathogen Disease-causing organism.

Peptide Two or more amino acids joined by a linkage called a peptide bond.

Pesticide Any substance intended for preventing, destroying, attracting, repelling, or controlling any pest including unwanted species of plants or animals during the production, storage, transport, distribution and processing of food, agricultural commodities, or animal feeds, or which may be administered to animals for the control of ectoparasites. The term includes substances intended for use as a plant-growth regulator, defoliant, desiccant, fruit-thinning agent, or sprouting inhibitor, and substances applied to crops either before or after harvest to protect the commodity from deterioration during storage and transport. The term normally excludes fertilizers, plant and animal nutrients, food additives, and animal drugs.

Phenotype Observable characteristics, resulting from interaction between an organism’s genetic makeup and the environment. Cf. Genotype

Phenylpropanoids Especially the derivatives of the cinnamyl alcohols and of cinnamic acids, isolated from medicinal plants due to the interest as the source for the preparation of the remedies.

Photosynthesis Conversion by plants of light energy into chemical energy, which is then used to support the plants’ biological processes.

Phytate (phytic acid) A phosphorus-containing compound in the outer husks of cereal grains that, in addition to limiting the bioavailability of phosphorus itself, binds with minerals and inhibits their absorption.

Phytochemicals Small molecule chemicals unique to plants and plant products.

Plasmid Circular extrachromosomal DNA molecules present in bacteria and yeast. Plasmids replicate autonomously each time the organism, a bacterium, divides and are transmitted to the daughter cells. DNA segments are commonly cloned using plasmid vectors.

Plasticity The quality of being plastic or able to be molded, changed.

Plastid Any of various small particles in the cytoplasm of the cells of plants and some animals that contain pigments (see chromoplast), starch, oil, protein, etc.

Pleiotropic Genes or mutations that result in the production of multiple effects at the phenotypic level. It is the consequence of the fact that biochemical pathways starting from different genes intersect in many places, inhibiting, deflecting, and variously modifying each other. Introduced genes may also insert into sites that effect phenotypic changes other than the one desired.

Polyclonal Derived from different types of cells.

Polymer A long molecule of repeated subunits.

Polypeptide Long chain of amino acids joined by peptide bonds. *Post-Transcriptional Gene Silencing* Post-transcriptional gene silencing (PTGS) is a sequence-specific RNA degradation system designed to act as an antiviral defense mechanism. A form of PTGS triggered by transgenic DNA, called co-suppression, was initially described in plants and a related phenomenon, termed quelling, was later observed in the filamentous fungus *Neurospora crassa*.

Posttranscriptional modification A series of processes through which protein molecules are biochemically modified within a cell following their synthesis by translation of messenger RNA. A protein may undergo a complex series of modifications in different cellular compartments before its final functional form is produced.

Profiling Creation of indiscriminate patterns of the substances within a sample with the aid of analytical techniques, such as functional genomics, proteomics, and metabolomics. The identity of the compounds detectable within the pattern need not be known.

Promoter A DNA sequence that is located near or even partially within encoding nucleotide sequences and which controls gene expression. Promoters are required for binding of RNA polymerase to initiate transcription.

Protein Proteins are biological effector molecules encoded by an organism's genome. A protein consists of one or more polypeptide chains of amino

acid subunits. The functional action of a protein depends on its three-dimensional structure, which is determined by its amino acid composition and any posttranscriptional modifications.

Proteomics The development and application of techniques used to investigate the protein products of the genome and how they interact to determine biological functions. This is an "Open ended" analytical technique that generates profiles of the proteins within a biological sample. The technique is commonly used to find differences between profiles of different (groups of) samples and to determine the identity of the associated proteins elucidated. Contrary to targeted analysis, these techniques are indiscriminate in that they do not require prior knowledge of every single substance protein present that is analyzed beforehand.

Protoplast fusion The fusion of two plant protoplasts that each consists of the living parts of a cell, including the protoplasm and cell membrane but not the vacuoles or the cell wall.

Protoplast The cellular material that remains after the cell wall has been removed. A plant cell from which the cell wall has been removed by mechanical or enzymatic means. Protoplasts can be prepared from primary tissues of most plant organs as well as from cultured plant cells.

Quantitative trait loci The locations of genes that together govern a multigenic trait, such as yield or fruit mass.

Recombinant DNA Any DNA molecule formed by joining DNA segments from different sources (not necessarily different organisms). This may also be a strand of DNA synthesized in the laboratory by splicing together selected parts of DNA strands from different organic species, or by adding a selected part to an existing DNA strand.

Regeneration Laboratory technique for forming a new plant from a clump of plant cells.

Regulatory gene A gene that acts to control the protein-synthesizing activity of other genes.

Regulatory sequence A DNA sequence to which specific proteins bind to activate or repress the expression of a gene.

Replication Reproduction or duplication, as of an exact copy of a strand of DNA.

- Rhizobium** A class of microorganisms that converts atmospheric nitrogen into a form that plants can utilize for growth. Species of this microorganism grow symbiotically on the roots of certain legumes such as peas, beans, and alfalfa.
- Ribonucleic acid (RNA)** A molecule similar to DNA that functions primarily to decode the instructions for protein synthesis that are carried by genes. (*See also* Messenger RNA Transfer RNA).
- Ribosome** A cellular component, containing protein and RNA, that is involved in protein synthesis.
- Ribozyme** Any of the RNA molecules possessing catalytic activity and acting as biological catalysts.
- Risk** A function of the probability of an adverse health effect and the severity of that effect, consequential to a hazard(s).
- Risk analysis** A process consisting of three components: risk assessment, risk management, and risk communication.
- Risk assessment** A scientific process consisting of the following steps: (1) hazard identification, (2) hazard characterization, (3) exposure assessment, and (4) risk characterization.
- Risk characterization** The qualitative and/or quantitative estimation, including attendant uncertainties, of the probability of occurrence and severity of known or potential adverse health effects in a given population based on hazard identification, hazard characterization, and exposure assessment.
- Risk communication** The interactive exchange of information and opinions throughout the risk analysis process concerning hazards and risks, risk-related factors, and risk perceptions, among risk assessors, risk managers, population, industry, the academic community, and other parties, including the explanation of risk assessment findings and the basis of risk management decisions.
- Risk management** The process, distinct from risk assessment, of weighing policy alternatives, in consultation with all interested parties, considering risk assessment and other factors relevant for the health protection of population and for the promotion of fair practices, and if needed, selecting appropriate prevention and control options.
- RNAi** RNA Interference (RNAi), a term coined by Fire et al. in 1998, is a phenomenon whereby small double-stranded RNA (referred as small interference RNA or siRNA) can induce efficient sequence-specific silence of gene expression.
- SAFOTEST** EU project on new methods for the safety testing of transgenic food.
- Scale-up** Transition from small-scale production to production of large industrial quantities.
- Secondary metabolites** Chemical substances within a biological organism sample that are not necessary for, or concerned with, primary cellular functions. Examples of secondary metabolites are Secondary metabolism proceeds by modification of the primary metabolites of photosynthesis, respiration, etc., by four main pathways. The malonate/polyketide pathway leads to the production of fatty acids and naphthoquinones. The mevalonate/isoprenoid pathway leads to the various terpenes (such as menthol), carotenoids and steroids. The shikimate pathway leads to aromatic amino acids and the phenolics and the final group of metabolites is a nonspecific mix of amino-acid derivatives including the and alkaloids (such as solanine) and others of mixed biogenesis.
- Selectable marker** A gene, often encoding resistance to an antibiotic or an herbicide, introduced into a group of cells to allow identification of those cells that contain the gene of interest from the cells that do not. Selectable markers are used in genetic engineering to facilitate identification of cells that have incorporated another desirable trait that is not easy to identify in individual cells.
- Selective breeding** Making deliberate crosses or matings of organisms so the offspring will have particular desired characteristics derived from one or both of the parents.
- Selective medium** Nutrient material constituted such that it will support the growth of specific organisms while inhibiting the growth of others.
- Sequence homology** The measurable likeness degree of identity or similarity between two nucleotides or amino acid sequences.
- Sera-binding tests** Immunological assays that evaluate for the presence of antigen-specific IgE in blood serum obtained from individuals allergic to food,

pollen, or other environmental antigens. Sera-binding tests include assays such as western blotting, ELISA, ELISA-inhibition, RAST, and RAST-inhibition techniques.

Shikimate pathway Pathway in microorganisms and plants involved in the biosynthesis of the aromatic amino acid family (phenylalanine, tyrosine, tryptophan) with a requirement for chorismate as well as shikimate. Secondary metabolites such as lignin, pigments, UV light protectants, phenolic redox molecules, and other aromatic compounds, such as folic acid and ubiquinone, are postscript products of the shikimate pathway.

Signal transduction The molecular pathways mechanism through which a cell senses changes in its external environment and changes its gene expression patterns in response.

Signal sequence The N-terminal sequence of a secreted protein, which is required for transport through the cell membrane.

Small interfering RNA (siRNA) Small Interfering RNA (siRNA) is 21–23-nt double-stranded RNA molecules. It guides the cleavage and degradation of its cognate RNA.

Site-specific recombination A crossover event, such as the integration of phage lambda, that requires homology of only a very short region and uses an enzyme specific for that recombination. Recombination occurring between two specific sequences that need not be homologous, mediated by a specific recombination system.

Somaclonal selection Epigenetic or genetic changes, sometimes expressed as a new trait, resulting from in vitro culture of higher plant cells. Somatic (vegetative nonsexual) plant cells can be propagated in vitro in an appropriate nutrient medium. The cells which multiply by division of the parent somatic cells are called somaclones and, theoretically, should be genetically identical with the parent. In fact this process occasionally in vitro cell culture of somatic cells, whether from a leaf, a stem, a root, a shoot, or a cotyledon, frequently generates cell plants which are significantly different, epigenetically and/or genetically, from the parent in a stable fashion and such progenies are called somaclonal variants and may provide a useful source of genetic variation.

Stilbenes A colorless or slightly yellow crystalline water-insoluble unsaturated hydrocarbon used in the manufacture of dyes trans-1,2-diphenylethene. Formula: $C_6H_5CH:CHC_6H_5$. It forms the backbone structure of several compounds with estrogenic activity. Trans-3,4',5-trihydroxy-stilbene also known as resveratrol, has been found in some experiments to inhibit cell mutations, and stimulate at least one enzyme that can inactivate certain carcinogens, and may contribute to a low incidence of cardiovascular disease.

Structural gene A gene that codes for a protein, such as an enzyme.

Substantial equivalence In the report of the 1996 FAO/WHO Expert Consultation, substantial equivalence was identified as being “established by a demonstration that the characteristics assessed for the genetically modified organism, or the specific food product derived therefrom, are equivalent to the same characteristics of the conventional comparator. The levels and variation for characteristics in the genetically modified organism must be within the natural range of variation for those characteristics considered in the comparator and be based upon an appropriate analysis of data.” In the Codex Guideline for the Conduct of Food Safety Assessment of Foods Derived from Recombinant-DNA Plants (2003), the concept of substantial equivalence is described as “a key step in the safety assessment process. However, it is not a safety assessment in itself rather it represents the starting point which is used to structure the safety assessment of a new food relative to its conventional counterpart. This concept is used to identify similarities and differences between the new food and its conventional counterpart. It aids in the identification of potential safety and nutritional issues and is considered the most appropriate strategy to date for safety assessment of foods derived from recombinant-DNA plants. The safety assessment carried out in this way does not imply absolute safety of the new product; rather it focuses on assessing the safety of any identified differences so that the safety of the new product can be considered relative to its conventional counterpart.”

Substrate Material acted on by an enzyme.

Synteny All loci on one chromosome are said to be syntenic (literally on the same ribbon). Loci may appear to be unlinked by conventional genetic tests for linkage but still be syntenic.

Systems biology A biology-based interdisciplinary study field that focuses on complex interactions in biological systems claiming that it uses a new perspective (holism instead of reduction). Particularly, from year 2000 onward, the term is used widely in the biosciences, and in a variety of contexts. An often stated ambition of systems biology is the modeling and discovery of emergent properties, properties of a system whose theoretical description is only possible using techniques which fall under the remit of systems biology.

Tannins Any of a class of yellowish or brownish solid compounds found in many plants and used as tanning agents, mordants, medical astringents, etc. Tannins are derivatives of gallic acid with the empirical formula $C_{76}H_{52}O_{46}$.

T-DNA The segment of the Ti plasmid of *A. tumefaciens* that is transferred to the plant genome following natural infection.

Ti Plasmid A plasmid containing the gene(s) responsible for inducing plant tumor formation transfer of genes from *A. tumefaciens* to plant cells.

Tissue culture In vitro growth in nutrient medium of cells isolated from tissue.

Traditional breeding Modification of plants and animals through selective breeding. Practices used in traditional plant breeding may include aspects of biotechnology such as tissue culture and mutational breeding.

Transcription The process through which a gene is expressed to generate a complementary messenger RNA molecule. Synthesis of messenger (or any other) RNA on a DNA template.

Transcription activator–like effector nucleases (TALENs) Transcription activator–like effector (TALE) proteins from *Xanthomonas* are nucleases that cleave unique genomic sequences in living cells and can be used for targeted gene editing and mutagenesis.

Transcriptome The total messenger RNA expressed in a cell or tissue at a given point in time.

Transgene A gene from one source that has been incorporated into the genome of another organism.

Transgenic plant A fertile plant that carries an introduced gene(s) in its germ line.

Transformation Change in the genetic structure of an organism by the incorporation of foreign DNA.

Transgenic organism An organism formed by the insertion of foreign genetic material into the germ line cells of organisms. Recombinant DNA techniques are commonly used to produce transgenic organisms.

Translation Process by which the information on a messenger RNA molecule is used to direct the synthesis of a protein.

Transmissible spongiform encephalopathy A disease that can be transmitted from one animal to another and will produce changes in the brain that appear similar to a sponge (i.e., some of the cells are clear when seen down the microscope).

Transposon A segment of DNA that can move around and be inserted at several sites in the genome of a cell possibly altering expression. The first to be described was the Ac/Ds system in maize shown by McClintock to cause unstable mutations.

Trypsin inhibitors Antinutrient proteins present in plants such as soybeans that inhibit the digestive enzyme, trypsin if not inactivated by heating or other processing methods.

Unintended effect An effect that was not the purpose of the genetic modification or mutation. An unintended effect may be either predictable or unpredictable, based on the knowledge of, among other things, the function of the introduced DNA and of the native DNA affected by the genetic modification. A predicted unintended effect would be, for example, variations in metabolic intermediates and endpoints, and an unpredicted effect might be turning on of unknown endogenous genes.

Variety A subdivision of a species for taxonomic classification also referred to as a “cultivar.” A variety is a group of individual plants that is uniform, stable, and distinct genetically from other groups of individuals in the same species.

Virulence Ability to infect or cause disease.

Virus A submicroscopic organism that contains genetic information but cannot reproduce by itself. To replicate, it must invade another cell and use parts of that cell's reproductive machinery.

Wildtype The form of an organism that occurs most frequently in nature.

Definition of the Subject

New and innovative techniques will be required to improve the efficiency of the global agriculture sector to ensure an ample supply of healthy food. To confront this situation the inequity between the affluent and developing countries will continue to grow and only a handful of technologies are sufficiently scale neutral to help with redressing this imbalance. Biotechnology is one such technology which offers efficient and cost-effective means to produce a diverse array of novel, value-added products and tools. The first generation of biotechnology products commercialized were crops focusing largely on input agronomic traits whose value was often opaque to consumers. The coming generations of crop plants can be grouped into four broad areas each presenting what, on the surface, may appear as unique challenges and opportunities. The present and future focus is on continuing improvement of agronomic traits such as yield and abiotic stress resistance in addition to the biotic stress tolerance of the present generation; crop plants as biomass feedstocks for biofuels and "bio-synthetics"; value-added output traits such as improved nutrition and food functionality; and plants as production factories for therapeutics and industrial products. From a consumer perspective the focus on value-added traits, especially improved nutrition, is undoubtedly one of the areas of greatest interest.

Introduction

During the coming decades, food and agricultural production systems will need to be significantly enhanced to respond to a number of remarkable changes, such as a growing world population; increasing international competition; globalization; shifts to increased meat consumption in developing countries and rising consumer demands for improved food quality, safety, health enhancement, and convenience. The 2008

World Bank Development Report emphasized that "Agriculture is a vital development tool for achieving the Millennium Development Goals that call for halving by 2015 the share of people suffering from extreme poverty and hunger [1]." The Report notes that three out of every four people in developing countries live in rural areas and most of them depend directly or indirectly on agriculture for their livelihoods. It recognizes that overcoming abject poverty cannot be achieved in sub-Saharan Africa without a revolution in agricultural productivity for resource-poor farmers in Africa, many of whom are women.

New and innovative techniques will be required to ensure that this revolution produces an ample supply of nutritious food by improving the efficiency of the global agriculture sector. Innovation is essential for sustaining and enhancing agricultural productivity. This involves new, science-based products and processes that contribute reliable methods for improving quality, productivity, and environmental sustainability. Biotechnology has introduced a new dimension to such innovation, offering efficient and cost-effective means to produce a diverse array of novel, value-added products and tools. It has the potential to improve qualitative and quantitative aspects of food, feed, fiber, and biofuel production; reduce the dependency of agriculture on chemicals and fossil fuels; diminish overcultivation and erosion; and lower the cost of raw materials, all in an environmentally sustainable manner. Commercialization of the first generation of products of recombinant DNA technology was another facet in a long history of human intervention in nature for agricultural and food production purposes. As such, the same parameters of risk-based assessment should apply. Commercialization of products must be undertaken within a regulatory framework that ensures adequate protection of the consumer, the environment, and alternate production systems while not stymieing innovation.

In a world whose population is increasing disproportionately in disadvantaged regions, it is hard to envisage feeding and sustaining these numbers in a livable environment without the use of biotechnology. From 1800 onward, more food was simply produced by plowing up virgin land and forest. The land area used for farming increased about fivefold up to the middle of the twentieth century in step with population

increases. The Green Revolution put a brake on this expansion, increasing yields threefold with limited need for further expansion. Since 1950, the proportion of the land devoted to farming has barely increased, even though the world population doubled over the same period. At least half the available good quality soil is currently used for agriculture, with the remainder under tropical forests. Coupling this with the ever diminishing nonrenewable resources and the compounding effects of climate change on the limitation of land usage leads to an obvious dilemma. Unless a second Green Revolution is carried out, increasing yield but limiting it to land currently used for farming, there will be further deterioration of natural habitats and biodiversity that may threaten more than our lifestyles.

During the most recent global food crisis in 2008 which was erroneously laid disproportionately on the shoulders of biofuel production, most especially grain ethanol, the Gates Foundation announced \$306 million in grants to boost agricultural yields in the developing world, with nearly \$165 million to replenish depleted soils in Africa. As noted by US News and World reports these efforts are not without controversy as they charge that critics consider that western philanthropists are violating African “food sovereignty” and promoting American at the expense of peasant farmers who are knowledgeable about local practices [2]. But local practices have yielded scarcity. A farmer in India grows three to four times as much food on the same amount of land as a farmer in Africa; a farmer in China, roughly seven times as much.

As noted, the FAO reports that global demand for food could easily double over the period 2000–2050, with a two-and-a-half- threefold increase in the poorest countries [2]. They found that biotechnology and genetic engineering of crops hold great promise for agriculture in developing countries. The report noted that more than 70% of the world’s poor still live in rural areas and depend directly on agriculture for their survival. The WHO estimates that 800 million people worldwide suffer from malnutrition. It is difficult to imagine a promising alternative to biotechnology and industrial agriculture that will sustain such numbers without catastrophic consequences. As far back as 2004 the Economic and Social Council of the United Nations (ECOSOC) noted that most developing countries are

unlikely to meet the Millennium Development Goals without a clear political commitment to making science and technology among top priorities in their development agenda [3]. FAO members called for strengthening efforts in maximizing the benefits and minimizing the potential adverse consequences of biotechnology, through the Committee on Agriculture, the Council and the Conference, the development of a multidisciplinary, cross-sectoral program. In response, the Biotechnology Applications in Food and Agriculture, Forestry and Fisheries Priority Area for Interdisciplinary Action (Biotech-PAIA) was established and an Inter-Departmental Working Group on Biotechnology was set up to oversee its planning and implementation [4]. And prior to that the US National Academy of Sciences, joined by six other academies from around the world (Royal Society of London, Third World Academy of Sciences, and National Academies of Brazil, China, India, and Mexico) issued a report in 2000 declaring that biotechnology should be used to increase the production of main food staples, improve the efficiency of production, reduce the environmental impact of agriculture, and provide access to food for small-scale farmers [5]. Agricultural research of all forms holds an important key to meeting their needs, as the FAO noted biotechnology can speed up conventional breeding programs and may offer solutions where conventional methods fail. This is a positive outcome for consumers and the environment.

Progress to Date

Modifications of crop plants can be organized into two broad-based non-mutually exclusive categories: those that benefit the producer and those that benefit the consumer. Modifications that protect the crop from either biotic or abiotic stress (biotic stress being damage by predators such as insects and nematodes and disease agents such as viruses, fungi, bacteria, and weeds, and abiotic stress in the form of drought, cold, heat, and poor soils), or increase in total crop yield benefit the producer and are called “input traits.” The majority of modified crops in commercial use fit in this group. Scientists have just begun to tap the large potential of biotechnology to produce varieties of plants that confer a wide spectrum of advantages to consumers. These varieties are modified with “output traits.”

Developing and commercializing plants with these improved traits involves overcoming a variety of technical, regulatory, and perception challenges inherent in perceived and real challenges of complex modifications. Both the panoply of traditional plant breeding tools and modern biotechnology-based techniques will be required to produce plants with the desired quality traits. In addition to the older gene transfer technology where mostly single genes were modified, newer techniques such as the use of RNA interference to manipulate endogenous genes and especially the use of transcription factors to modulate whole suites of genes and metabolic networks will become increasingly important tools in the effort to introduce valuable traits. The later approach is already a major focus in multigenic and quantitative traits such as developing stress tolerance crops and modifying paths for improving nutritional characteristics.

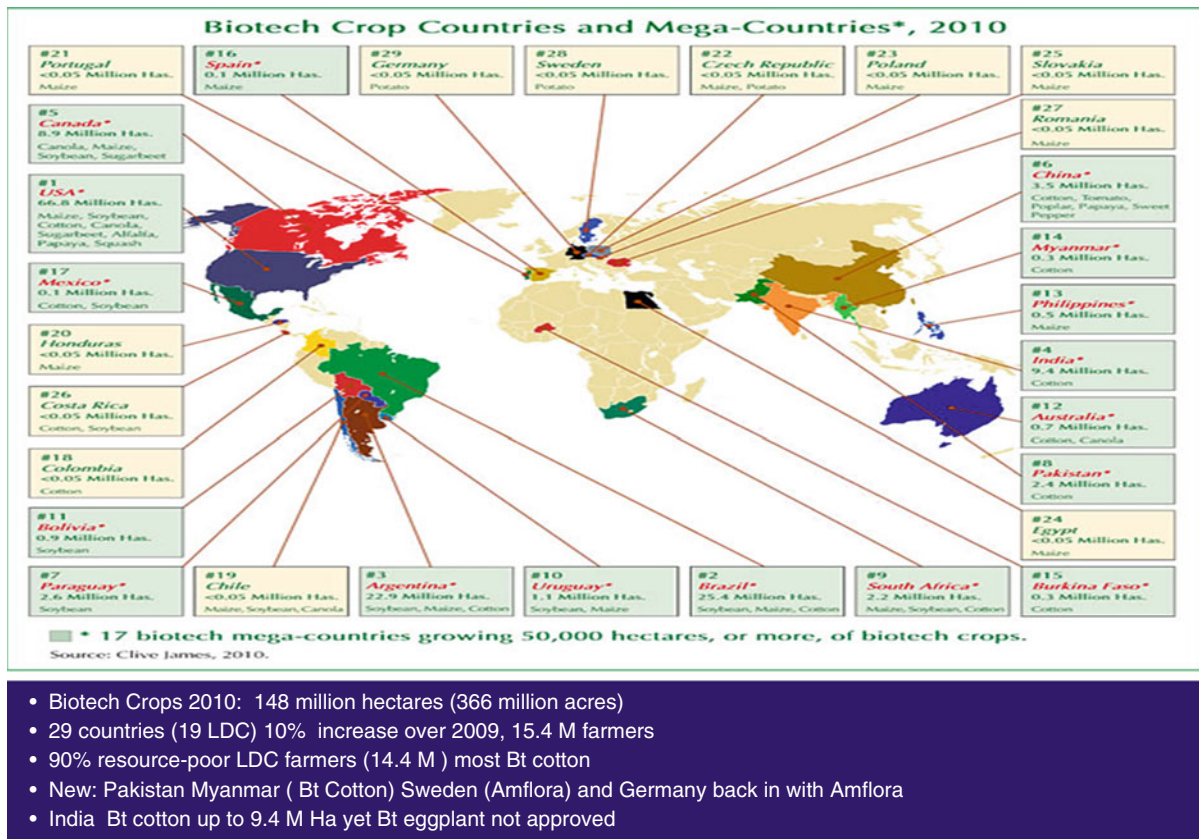
Since the first biotech crop was commercialized in 1996, genetically modified (GM) crops are now grown commercially by 15.4 million farmers in 29 countries on 366 million acres [6] (Fig. 1). More than half of the 63 countries engaged in biotech research, development, and production are developing countries. While North America still leads with US acreage accounting for about 45% of the total acreage worldwide, nevertheless, 19 of the 29 countries are developing countries and of the 15.4 million farmers that grew these crops, a full 14.4 million (90%) are resource-poor LDC farmers. The most recent countries to join this group were, in 2009, South America (Uruguay, Paraguay, and Bolivia) and Africa (Egypt and Burkina Faso) and in 2010, three countries planted approved biotech crops for the first time and Germany resumed planting. Pakistan planted Bt cotton, as did Myanmar, and notably Sweden, the first Scandinavian country to plant a biotech crop, planted “Amflora,” a potato with high amylase starch for industrial applications. Germany also resumed adoption of biotech crops by planting “Amflora” for a net gain of four countries in 2010 [6]. The first generation of such crops focused largely on input agronomic traits, and the next generation will focus more on value-added output traits. In the next decade, some studies estimate, the global value of biotech crops will increase nearly fivefold to \$210 billion [7].

Agricultural biotechnology has helped farmers around the world boost their productivity and grow

crops in more ecologically healthy fields while allowing much more efficient use of resources. This technology allows reduced tillage, which cuts down on greenhouse gas emissions, water runoff, machinery use, and soil erosion. Meanwhile, the benefits experienced by larger-scale farmers in both industrialized nations and lesser-developed countries are already considerable [6]. Research by Brookes and Barfoot [8, 9] shows in the first 11 years of GM crop cultivation that global net farm income increased by \$33.8 billion since 1996; the environmental footprint associated with pesticide use was reduced by 15.4%; and there was a reduction in carbon dioxide emissions in 2006 equivalent to taking nearly 6.6 million cars off the road for a year.

An earlier study by researchers at Denmark’s National Environmental Research Institute (NERI) monitored the fields of conventional and glyphosate-tolerant sugar beet. They found that the GM plots supported more plant species and insects than the conventional plots, thus providing more food for birds, and other types of wildlife use of transgenic crops increased biodiversity compared to traditional herbicide treatments [10]. Proper measurements in the UK indicate that no-till, (directly compared with plowed organic fields on the same farm and using the same farmer) uses only one third fossil fuel, uses land much more efficiently, reduces nitrate (and pesticide) runoff by at least half, and increases soil carbon which is lost when plowed. In addition, bird territories are orders of magnitude higher, soil erosion almost vanishes, and soil invertebrates such as earthworms soar in numbers, as do predatory arthropods to keep pests down. Organic fields in the UK see a threefold rise in weeds on conversion that necessitates use of the plow [11].

Therefore, reduced-till agriculture means healthier soil, with reduced erosion and far less carbon dioxide release. Soil carbon sequestration will be an important part of any international strategy to mitigate the increase in atmospheric CO₂ concentrations. By adopting more sustainable management practices, agriculture may play a large part in enhancing soil carbon sequestration across the globe. One way is by reducing the amount of conventional tillage, after long-term tillage soil carbon stocks are depleted. In general, cultivation is not a sustainable practice. It is energy intensive and exposes soil to wind and water erosion.



Transgenic Crops, Next Generation. Figure 1

Global map of biotech crop countries and mega-countries in 2010

It allows rain to compact the soil and increases the oxygen content thus allowing organic matter to oxidize away. In turn, lower organic matter in the soil allows more compaction and more nutrient loss. Additionally in warmer and drier climates, evaporative water loss may be reduced as residue remains on the soil surface creating a wetter and cooler soil microclimate.

The Brooks and Barfoot study indicates that pesticide use fell by over 286 million kg (−7.8%: equivalent to about 40% of the annual volume of pesticide-active ingredient applied to arable crops in the European Union). Less spraying means fewer tractor passes, contributing to lower carbon dioxide emissions. Insect-resistant maize also has a collateral effect – less insect damage results in much less infection by fungal molds which reduces mycotoxins that are known health risks causing such problems as liver

cancer to humans and animals. Bt corn resulted in a 90% reduction in mycotoxin fungal fumonisins [12]. In addition to the obvious health benefits, the total US economic benefit is estimated to be approximately \$23million annually [6]. The only “natural” way to control those fungi is the use of copper sulfate which has one of the highest toxic hazard ratings of acceptable pesticides and selects for antibiotic-resistant bacteria in the soil.

A 2005 paper from the Royal Society suggests that intensive high-yield farming on less land is better for wildlife than “wildlife-friendly” less-efficient farming [13]. They provide convincing evidence that without yield increase, land use will double by 2050 and that this effect will be especially significant in developing countries where, without greater productivity China and India will need four times the land area to support

their expanding populations. They show that in Latin America, where increased productivity was achieved, there was a significant decrease in deforestation; those producers with greatest yield increase had lower land use.

While North America remains the epicenter for cutting edge GM research, other regions, namely, China are emerging as contenders on the global stage. Agricultural science is now China's fastest-growing research field with China's share of global publications in agricultural science growing from 1.5% in 1999 to 5% in 2008 [6]. China's early experience with Bt cotton demonstrated the direct and indirect benefits of its investment in plant biotechnology research and product development. In 2002, Bt cotton was grown in 2.1 million hectares by around five million farmers. At that time the average Bt cotton farmer had reduced pesticide sprayings for the Asian bollworm from 20 to 6 times per year, reduced applications by 59–80% compared to conventional cotton (assessed in 3 years of use), and produced a kilogram of cotton for 28% less cost than the farmer using non-Bt varieties. Net revenues increased by 357–549 USD/ha compared to conventional cotton (assessed in 3 years of use) [14]. Ultimately, however, it is the social benefits from reducing exposure to insecticides and saving lives which is the real payoff.

The demand for productivity-enhancing technologies by farmers and for cost savings by consumers, the rate of increase in research investments, and success with Bt cotton suggest that products from China's research program will 1 day become widespread inside China. Indeed China is emerging as one of the trend-setters in the adoption of novel traits as more recently China is setting the pace for new approvals, with Bt rice and phytase maize approved on November 27, 2009. Rice is the principal staple for much of the world and maize is the largest animal feed source. Bt rice has the potential to increase yields up to 8%, decrease pesticide use by 80% (17 kg/ha), and generate US\$4 billion in benefits annually [6]. The phytase approval is a major step forward in approvals as it is the first since the FLAVR SAVR tomato focusing on a "quality" trait. However, it is far more than this, both literally and figuratively since this single trait addresses several issues from nutritional to environmental as expanded on later.

The first GM crop to be released for commercial cultivation in India was Bt cotton, developed by the Maharashtra Seed Company (Mahyco) in partnership with Monsanto. The approval, which was given in 2002, came after several years of field trials following the biosafety procedures laid down by the government. Three cotton hybrids were granted permission for field sowing in six states for 3 years. For the first season, farmer demand for Bt cotton seed was very high; it is estimated that 44,500 ha of certified Bt cotton were planted by nearly 55,000 farmers. However, the initial events thrived in regions that resembled the area in which they were originally developed but did not perform as well in growing regions with disparate climate challenges. It was not until the trait was introgressed into locally adapted varieties that Bt cotton thrived in all growing regions. Between 2005 and 2006 the biggest impact of this approach was realized. From 3 Bt cotton hybrids in 2002 to 62 in 2006 the rapid deployment of Bt cotton hybrids based on different agro-climatic conditions resulted in decreased insecticide sprays by 39%, and increased yield of 31%, resulting in increased profit per hectare of 88% or \$ 250. Over this period of rapid deployment the average cotton yields increased from 308 to 450 kg/ha of lint (of this increase 50% could be attributed to Bt technology). Over the same period raw cotton exports rose from 0.9 million bales in 2005 to 4.7 million in 2006 and had achieved 5.9 million by 2007 [15]. By 2009, 5.6 million resource-poor farmers in India planted 8.4 million hectares of Bt cotton, equivalent to 87% of the 9.6 million hectare national cotton crop. The increase from 50,000 ha when Bt cotton was first commercialized in 2002 to 8.4 million hectares in 2009 represents an unprecedented 168-fold increase in 8 years. Between 2002 and 2008, Bt cotton generated economic benefits for farmers valued at US \$5.1 billion, halved insecticide requirements, contributed to the doubling of yield and transformed India from a cotton importer to a major exporter. Choudary contends that the deployment of Bt cotton over the last 8 years has resulted in India becoming the number one exporter of cotton globally as well as the second largest cotton producer in the world [15].

However, despite the success of Bt cotton the expected successful commercialization of Bt eggplant never materialized as an effective opposition managed to scupper its approval. Bt eggplant, or brinjal as it is

referred to in India, was found to be effective against fruit and shoot borer (FSB), with 98% insect mortality in shoots and 100% in fruits compared to less than 30% mortality in non-Bt counterparts. The multi-location research trials confirmed that Bt brinjal required, on average, 77% less insecticides than non-Bt counterparts for control of FSB, and 42% less for the control of all insect pests of brinjal. The benefits of Bt brinjal translate to an average increase of 116% in marketable fruits over conventional hybrids, and 166% increase over popular open-pollinated varieties [16]. Furthermore, the significant decrease in insecticide usage reduced the farmers' exposure to insecticides and resulted in a substantial decline in pesticide residues on brinjal fruits. Scientists have estimated that Bt brinjal will deliver farmers a net economic benefit ranging from \$330 to \$397 per acre with national benefits to India exceeding \$400 million per year. However in February, 2010, the environmental minister announced a 6-month moratorium citing that "There is no overriding food security argument for Bt brinjal. Our objective is to restore public confidence and trust in Bt brinjal" [17], clearly articulating the fact that the decision was not based on scientific analysis or risk assessment.

A number of other multi-institutional projects have also been launched, including the development of transgenic plants for resistance to geminiviruses in cotton, mungbean, and tomato, resistance to rice tungro disease, development of a nutritionally enhanced potato with a balanced amino acid composition, and development of molecular methods for heterosis breeding. Other transgenic crops that are awaiting approval for commercial cultivation include transgenic herbicide-tolerant mustard hybrids and nutritionally enhanced potato varieties. However, despite the resounding success of Bt cotton, given the experience with Bt brinjal it is difficult to be optimistic about the prospects for commercialization of food crops.

A somewhat similar but even more insidious situation was experienced by Egypt. In 2008, Egypt became the first country in the Arab world (and only one of three in Africa) to commercialize biotech crops by planting 700 ha of Bt yellow maize. The variety commercialized, Ajeeb-YG, is a cross between MON 810 and an Egyptian maize variety with resistance to three

corn borer pests. The reason that the amount was so low can in part be attributed to political pressure from their EU market.

The Next Generation

The vast majority of products approved to date are in the area of agronomic traits, most specifically biotic stress. The principal focus in the immediate future will remain on agronomic traits especially the area of pest control but with an increasing interest in abiotic stress tolerance which is gaining prominence as external pressures from climate change to land use change.

On the biotic stress tolerance side the focus is expanding to multi-tiered control systems. This in theory serves a double advantage, primarily expanding the effectiveness of the broad-based resistance events but also allowing more effective management of the resistance trait since there is less selective pressure when genes are stacked. SmartStax, an eight-trait event developed through collaboration between Monsanto and Dow takes advantage of multiple modes of insect protection and herbicide tolerance against above and below ground insects and provides broad herbicide tolerance, including Yieldgard VT Triple (Monsanto), Herculex Xtra (Dow), Roundup Ready 2 (Monsanto), and Liberty Link (Dow). It is currently available for corn, but cotton, soybean, and specialty crop variations are to be released [18]. It is estimated that this should require only 5% refuge acres as opposed to the 20% mandated for older technologies to mitigate against pest tolerance [15].

On the second area of agronomic traits, namely abiotic stress, there is a meta-issue that overlays much of the individual efforts, which is climate change. This poses a real challenge in terms of available agricultural land and freshwater use. Apart from the obvious effects of climate change, the decline of crop yields, ocean acidification, poor nutrition and abiotic stress, population displacement, and threatened ecosystems are effects underlined by the Stern Report [19] as potential consequences of climate change. In addition there are also broader, more systemic effects of drought beyond food insecurity such as decreased household income, the loss of assets due to slaughter of livestock, health threats due to the lack of water for hygiene and household uses, environmental degradation, and less-sustainable land management.

These effects should be considered in the light of growing population levels. In order to feed the overall population, the world will have to double its rate of agricultural production over the next 25 years, despite having already quadrupled it in the last 50 years. Severe drought accounts for half the world's food emergencies annually [1]. In 2003, the World Food Program spent US\$565 million in response to drought in sub-Saharan Africa (SSA). In this context, solutions must be developed to adapt crops to the existing but also evolving conditions, such as marginal soils or harsher conditions such as cold, heat, drought, and salinity. The agriculture sector is both a contributor and provider of potential solutions to this phenomenon. It impacts two of the principal components of climate change greenhouse gases and water. Agriculture is a major source of the former emissions. Practices – such as deforestation, cattle feedlots, and fertilizer use – currently account for about 25% of greenhouse gas emissions. When broken down this amounts to 14% of carbon dioxide emission, 48% of methane, and 52% of nitrous oxide emissions [19]. In addition, this sector uses a significant amount of available freshwater – approximately 70% of the water currently consumed by humans is used in agriculture – and this is likely to increase as temperatures rise.

Given the potential impacts of climate change on the range and extent of agricultural productivity and the impact of agriculture practices itself on global warming, techniques should play a substantial part in mitigating against climate change. Green biotechnology offers a set of tools which can help producers limit greenhouse gas emissions as well as adapting their agricultural techniques to shifting climates. The three major contributions of green biotechnology to the mitigation of the impact of climate change are greenhouse gas reduction, crop adaptation (environmental stress, changing niches) and protection, and yield increase in less desirable and marginal soils.

On the first of these issues greenhouse gas reduction in addition to carbon dioxide agriculture contributes two of the other major gases indeed one of them nitrous oxide has a global warming potential of about 300 times that of carbon dioxide. In addition, nitrous oxides stay in the atmosphere for a considerable period. Nitrous oxide is produced through bacterial degradation of applied nitrogen fertilizer. In addition, fertilizer

can contribute to eutrophication at ground level so its reduction is desirable on several levels. However, nitrogen is essential for crop production since it is quantitatively the most essential nutrient for plants and a major factor limiting crop productivity. One of the critical steps limiting the efficient use of nitrogen is the ability of plants to acquire it from applied fertilizer. Therefore, the development of crop plants that absorb and use nitrogen more efficiently can serve both the plant and the environment. Arcadia Biosciences of Davis, CA, developed nitrogen-efficient crops by introducing a barley AlaAT (alanine aminotransferase) into both rice and canola. Arcadia's Nitrogen Use Efficiency (NUE) technology produces plants with yields that are equivalent to conventional varieties but which require significantly less nitrogen fertilizer because the AlaAT gene allows more efficient use. Compared with controls, transgenic plants also demonstrated significant changes in key metabolites and total nitrogen content, confirming increased nitrogen uptake efficiency. This technology has the potential to reduce the amount of nitrogen fertilizer that is lost by farmers every year due to leaching into the air, soil, and waterways. In addition to environmental pressures, nitrogen costs can represent a significant portion of a farmer's input costs and can significantly impact farmer profitability. Farmers spend \$60 billion annually for 150 million tons of fertilizer [20]. The technology has been licensed to Dupont for maize and Monsanto for application in canola.

The second area where green technology can help in a changing climate is crop adaptation to environmental stress and changing niches. Under stress plants will divert energy into survival instead of producing biomass and reproduction, so addressing this impact should have substantial effect on yield. In addition, improved stress tolerance allows expanded growing season especially earlier planting and further reduces yield variability and grower financial risk. The most critical of these stresses is water. One of the most effective methods of addressing water limitation problems, namely, irrigation, unfortunately is also one of the major causes of arable land degradation. It is estimated that 24.7 million acres of farmland worldwide is lost each year due to salinity build up resulting from over irrigation. In fact crops are now limited by salinity on 40% of the world's irrigated land (25% of the USA).

To address the extreme end of irrigation impact Eduardo Blumwald at UC Davis used AtNHX1, the most abundant vacuolar Na⁺/H⁺ antiporter in *Arabidopsis thaliana* which mediates the transport of Na⁺ and K⁺ into the vacuole. By overexpressing this vacuolar Na⁺/H⁺ antiporter, transgenic tomatoes were able to grow, flower, and produce fruit in the presence of 200 mM sodium chloride [21]. Arcadia Biosciences had now introduced this gene into economically important crops.

Even at a more moderate level of impact it is estimated that about 70–80 million acres in USA suffer yield losses due to moderate water stress. The most critical time for water stress is near-pollination and flowering where yields with or without irrigation can vary by up to 100%. This effect is clearly demonstrable in dry land production where yields can be cut in half in the absence of irrigation. At this time about 15% of US maize acres are irrigated. Given the negative effective and cost of irrigation it is estimated 20 million acres in USA would benefit from a drought tolerance gene that gives a 10% yield increase. It would also allow shifting of high-value crops into production on more marginal land.

One of the first commercialized products to have included a “yield gene” is Monsanto’s second generation Roundup Ready 2 Yield[®] Soybeans which include not only the glyphosate-tolerant trait but also that which was developed using extensive gene mapping to identify specific DNA regions that segregated with yield increase. It is a perfect example of the power of combining recombinant DNA technology with genomics tools. The company claims that following 4 years of field trials across six US states showed 7–11% higher yields, compared to the first generation of Roundup Ready soybeans. At the National Technical Biosafety Committee (CTNBio) meeting in Brazil in August 2010, the committee approved the Bt enhanced version of this product for planting in Brazil [18].

As noted transcription factors are some of the most versatile tools being employed in developing stress-tolerant plants. One of the most versatile classes of transcription factors in so far as environmental response is concerned is the DREB (dehydration-responsive element binding protein) transcription factors which are involved in the biotic stress signaling pathway and can activate as many as 12 resistant functional genes relying on DREmembers of cis regulation under adverse conditions, for instance, rd29, cor15,

and rd17, cause proline content to rise so as to enable plants to improve in many resistances such as drought, freezing, and salt tolerance. It has been possible to engineer stress tolerance in transgenic plants by manipulating the expression of DREBs [22]. One isolated from *Arabidopsis* has improved drought tolerance increasing productivity by at least twofold during severe water stress. In Monsanto field trials using this approach, maize yields have increased under water stress by up to 30% [22].

Other approaches include modification of individual genes involved in stress response and cell signaling. For example, drought-tolerant canola engineered to reduce the levels of PARP [poly(ADP-ribose) polymerase], a key stress-related protein in many organisms, shows relative yield increases of up to +44% compared to control varieties. A subset of the transcription factors homeodomain leucine zipper proteins (HDZip) play a role in regulating adaptation responses including developmental adjustment to environmental cues such as water stress in plants [23]. One of these effectors is abscisic acid (ABA), an important plant regulator controlling many environmental responses including stomata movement which is itself modulated by the DREB elements. Some work is being done on modifying HDZip directly and others are working indirectly, for example, down regulating farnesyltransferase, a signaling system in the production of abscisic acid and stomata control, which results in stomata closure and water retention.

Investigators are also working on modifying basic acid to enhance the tolerance of plants to water-deficit by delaying the drought-induced leaf senescence and abscission during the stress episode. Using tobacco plants expressing an isopentenyltransferase (IPT) gene under the control of a stress- and maturation-induced promoter (PSARK) it was shown that delayed drought-induced leaf senescence resulted in remarkable drought-tolerant phenotypes, as well as minimal yield loss when plants were watered with only 30% of the water used under controlled conditions [24]. This is now being introduced into rice among other crops. This work is being done in conjunction with Arcadia Biosciences. In addition, Bayer CropScience, Pioneer Hi-Bred, BASF and Dow among others are conducting research on maize, cotton, canola, and rice, to develop a new generation of stress-tolerant, high-performance

crop varieties. Clearly stress-tolerant traits are of paramount importance in LDCs especially sub-Saharan Africa and Asia. Major efforts are already underway on this front. The partnership, known as Water Efficient Maize for Africa (WEMA), was formed in response to a growing call by African farmers, leaders, and scientists to address the devastating effects of drought on small-scale farmers. Frequent drought leads to crop failure, hunger, and poverty. Climate change can only aggravate this situation [25].

On the other end of the spectrum of climate change impact is flooding due to changing rain patterns and rising sea levels. This is already a major cause of rice crop loss. It is estimated that four million tons of rice are lost every year because of flooding which is sufficient to feed 30 million people. Rice is not grown in flooded fields through necessity but rather to control weeds, however, most rice varieties die after more than 3 days of complete submergence. Researchers know of at least one rice variety FR13A that can tolerate flooding for longer periods, but conventional breeding failed to create an event that was acceptable to farmers. The Ronald laboratory at UC Davis cloned the submergence tolerance (Sub1) locus from this resistance variety using a map-based cloning approach. The Sub1 locus encodes three putative transcription regulators one of which increases dramatically in response to oxygen deprivation in sub1 seedlings, whereas Sub1C levels decrease. Transgenic lines over-expressing the Sub1A-1 gene have been introgressed into a submergence-intolerant line and display-enhanced submergence tolerance [26].

There is also some research in the final abiotic stress focus area, namely, expansion of crops into and increased yield in less desirable and marginal soils. For example, a gene that produces citric acid in roots can protect plants from soils contaminated with aluminum as it binds to the contaminant preventing uptake by the root system [27]. Genes such as these can allow crops to be cultivated in hostile soils and temperatures increasing geographic range while reducing potential impact on fragile ecosystems.

While exciting and very relevant, research in abiotic stress tolerance is still an input trait. The first generation of biotechnology crops focused largely on those input agronomic traits; the next generation will focus more on value-added output

traits. This will include identifying and isolating genes and metabolites that will make possible the enhancement of valuable traits, with some of the later compounds being produced in mass quantities for niche markets. Two of the more promising markets are improved nutrition including nutraceuticals, or so-called functional foods, and plants developed as bioreactors (production factories) for the commercial-level production of valuable proteins and compounds, a field known as plant molecular farming [28]_ENREF_24.

While the correlative link between food and health, beyond meeting basic nutrition requirements, has only been unequivocally proven in a number of cases, a growing body of evidence indicates that food components can influence physiological processes at all stages of life [29]. Nutrition intervention from a functionality perspective has a personal dimension. Determining individual response is at least as complex a challenge as the task of increasing or decreasing the amount of a specific protein, fatty acid, or other components of the plant itself. There is also evidence that early food regimes can effect later life health, for example, some children that survived famine conditions in certain regions of Africa grew into adults battling obesity and related problems presumably due to the selective advantage of the thrifty gene in their early food-stressed environment becoming a hazard during more abundant times especially if later diets are calorie-dense.

Functional foods are defined as any modified food or food ingredient that may provide a health benefit beyond the traditional nutrients it contains. Scientific evidence is accumulating to support the role of phytochemicals and functional foods in the prevention and treatment of disease. Functional food components are of increasing interest in the prevention and/or treatment of a number of the leading causes of death including but not limited to cancer, diabetes, cardiovascular disease, and hypertension. Many food components are known to influence the expression of both structural genes and transcription factors in humans. Examples of these phytochemicals are listed in Table 2. The large diversity of phytochemicals suggests that the potential impact of phytochemicals and functional foods on human and animal health is worth examining as targets of biotechnology efforts. Developing plants

with improved quality traits involves overcoming a variety of technical challenges inherent to metabolic engineering programs. Both traditional plant breeding and biotechnology techniques are needed to produce plants carrying the desired quality traits [29] (Table 1).

From a health perspective, plant components of dietary interest can be broadly divided into four main categories, which can be further broken down into positive and negative attributions for human nutrition, macronutrients (proteins, carbohydrates, lipids [oils], and fiber), micronutrients (vitamins, minerals, phytochemicals), anti-nutrients (substances such as phytate that limit bioavailability of nutrients), allergens, intolerances, and toxins [29]. Developing and commercializing plants with these improved traits involves overcoming a variety of technical, regulatory, and perception challenges inherent in perceived and real challenges of complex modifications. Both the panoply of traditional plant breeding tools and modern biotechnology-based techniques will be required to produce plants with the desired quality traits. Table 2 presents examples of crops that have already been genetically modified with macro- and micronutrient traits that may provide nutritional benefits.

In addition to functional foods, this area has the potential to address both nutrition and environmental impact. A good example of this is the addition of transgenic phytase enzymes to crops to reduce the need to add phosphate to feed [6]. Most of the phosphate is added to counteract the non-bioavailability of phosphorus in phytic acid and the sequestering effect of phytic acid on uptake of divalent mineral ions such as iron, calcium, and zinc. Unfortunately excess phosphate is excreted causing major environmental impact through eutrophication and fish kills in regions with intense pig and poultry farming [28, 29]. In addition, in humans, such mineral deficiencies due to phytate binding are estimated to afflict two to three billion people, primarily in the developing world. Several studies have shown that *Aspergillus*-derived phytases can be produced in large quantities in a range of plants including cereals with clear-cut positive effects on phytate degradation, and phosphate and mineral bioavailability in animal-feeding trials [104]. It is thus conceivable that genetic engineering of staples for increased phytase expression could have potential for

improving iron and zinc bioavailability alleviating the need for supplementation in all monogastrics and consequent reduction in polluting runoff in non-ruminant animals [105]. As noted earlier China has led the way in the approval of this “output trait” in Maize being the first country to approve commercialization in November 2009 [6]. Continuing improvements in molecular and genomic technologies are contributing to the acceleration of such product development. One estimate states that foods that are used for functional purposes made up 10% of the \$503 billion total US retail food market [106].

In addition to being a source of nutrition, plants have been a valuable wellspring of therapeutics for centuries. During the past decade, however, intensive research has focused on expanding this source through rDNA biotechnology and essentially using plants and animals as living factories for the commercial production of vaccines, therapeutics, and other valuable products such as industrial enzymes and biosynthetic feedstocks [28].

More pressingly, with the increasing costs in economic and environmental terms of our dependency on fossil fuels, biotechnology offers innovative means to improve plant material for biomass conversion and enzymes to do the converting. There are two principal classes of biofuels: bio-alcohol (initially bio-ethanol but with increasing interest in higher-energy alcohols such as bio-butanol) and bio-diesel.

The first generation of biofuels was fermented from easy sources of simple sugars such as sugarcane and simple polymers primarily from grain starch. This source of bio-ethanol is unsustainable on many levels including the fact that its sources compete with food and especially feed grains for markets, land, and water. The focus for second generation bio-alcohols is mostly on complex polymers, primarily cellulosic ethanols and what are being termed third generation bio-alcohols such as bio-butanol [107]. From a biotechnology perspective work is being carried out on the biomass component focusing on increased production in such sources as switchgrass and miscanthus by among other things modifying photoperiodicity genes to switch energy to vegetative tissue production and improved biomass conversion by such approaches as reducing lignin composition and incorporating self-activating enzyme digestion upon harvesting. On the enzyme component companies including

Transgenic Crops, Next Generation. Table 1 Examples of crops in research with nutritionally improved traits^a

Trait	Crop (trait detail)	References
<i>Protein and amino acids</i>		
Protein quality and level	Bahiagrass (protein↑)	Luciani and Wofford [30]
	Canola (amino acid composition)	Roesler et al. [31]
	Maize (amino acid composition; protein↑)	Cromwell et al. 1969, [32], Yang et al. [33], O'Quinn et al. [34], Young et al. [35]
	Potato (amino acid composition; protein↑)	Chakraborty et al. [36], Li et al. [37], Yu and Ao [38], Atanassov et al. [39]
	Rice (protein↑; amino acid)	Katsube et al. [40]
	Soybean (amino acid balance)	Rapp [41]; Dinkins et al. [42]
	Sweet Potato (protein↑)	Prakash and Jaynes [43]
	Wheat (protein↑)	Uauy et al. [44]
Essential amino acids	Canola (lysine↑)	Falco et al. [45]
	Lupin (methionine↑)	White et al. [46]
	Maize (lysine↑; methionine↑)	Agbios 2006, Lai [47]
	Potato (methionine↑)	Zeh et al. [48]
	Sorghum (lysine↑)	Zhao et al. 2003
	Soybean (lysine↑; tryptophan↑)	Falco et al. [45], Galili et al. [49]
<i>Oils and fatty acids</i>		
	Canola (lauric acid↑; γ -linolenic acid↑; + ω -3 fatty acids; 8:0 and 10:0 fatty acids↑; lauric + myristic acid↑; oleic acid↑)	Del Vecchio [50], Froman [51], James et al. [52], Ursin [53], Dehesh et al. 1996, Agbios 2006, Roesler et al. [31]
	Cotton (oleic acid↑; oleic acid + stearic acid↑)	Chapman et al. [54], Liu et al. [55]
	Linseed (+ ω -3 and -6 fatty acids)	Abbadi et al. [56]
	Maize (oil↑)	Young et al. [35]
	Oil Palm (oleic acid↑ or stearic acid↑; oleic acid↑ + palmitic acid↓)	Parveez [57], Jalani et al. [58]
	Rice (α -linolenic acid↑)	Anai et al. [59]
	Soybean (oleic acid↑; γ -linolenic acid↑)	Kinney [60], Reddy and Thomas [61]
	Safflower (γ Linoleic Acid GLA↑)	Arcadia [62]
<i>Carbohydrates</i>		
Fructans	Chicory, (fructan↑; fructan modification)	Smeekens [63], Sprenger et al. [64], Sevenier et al. [65]
	Maize (fructan↑)	Caimi et al. [66]
	Potato (fructan↑)	Hellwege et al. [67]
	Sugar beet (fructan↑)	Smeekens [63]
Fructose, Raffinose, Stachyose	Soybean	Hartwig et al. [68]

Transgenic Crops, Next Generation. Table 1 (Continued)

Trait	Crop (trait detail)	References
Inulin	Potato (inulin↑)	Hellwege et al. [69]
Starch	Rice (amylase ↑)	Chiang et al. [70], Schwall et al. [71]
<i>Micronutrients and functional Metabolites</i>		
Vitamins and Carotenoids	Canola (vitamin E↑)	Shintani and DellaPenna [72]
	Maize (vitamin E↑; vitamin C↑; beta-carotene↑; folate↑)	Rochefford et al. [73], Cahoon et al. [74], Chen et al. [75], Naqvi et al. [76]
	Mustard (+β-carotene)	Shewmaker et al. [77]
	Potato (β-carotene and lutein↑)	Ducruex et al. [78]
	Rice (+β-carotene)	Ye et al. [79]
	Strawberry (vitamin C↑)	Agius et al. [80]
	Tomato (folate↑; phytoene and β-carotene↑; lycopene↑; provitamin A↑)	Della Penna 2007, Diaz de la Garza et al. [81], Enfissi et al. [82] Mehta et al. 2002, Fraser et al. [83], Rosati et al. [84]
Functional secondary metabolites	Apple (+stilbenes)	Szankowski et al. [85]
	Alfalfa (+resveratrol)	Hipskind and Paiva [86]
	Kiwi (+resveratrol)	Kobayashi et al. [87]
	Maize (flavonoids↑)	Yu et al. [88]
	Potato (anthocyanin and alkaloid glycoside↓; solanin↓)	Lukaszewicz et al. [89]
	Rice (flavonoids↑; +resveratrol)	Shin et al. 2006, Stark-Lorenzen et al. [90]
	Soybean (flavonoids↑)	Yu et al. [91]
	Tomato (+resveratrol; chlorogenic acid↑; flavonoids↑; stilbene↑anthocyanins↑)	Giovinazzo et al. [92], Niggeweg et al. [93], Muir et al. [94], Rosati et al. [84], Gonzali et al. [95]
	Wheat (caffeic and ferulic acids↑; +resveratrol)	UPI [96]
Mineral availabilities	Alfalfa (phytase↑)	Austin-Phillips et al. [97]
	Lettuce (iron↑)	Goto et al. [98]
	Rice (iron↑)	Lucca et al. [99]
	Maize(phytase↑, ferritin↑)	Drakakaki et al. [100], Han [101]
	Soybean (phytase↑)	Denbow et al. [102]
	Wheat (phytase↑)	Brinch-Pedersen et al. [103]

^aExcludes protein/starch functionality, shelf life, taste/aesthetics, fiber quality and allergen reduction traits. Modified from Newell-McGloughlin [29] (25)

Novozymes (Davis, CA) and Danisco (Palo Alto, CA) are making considerable strides in improving the effectiveness, specificity, and cost of cellulosic enzymes and increasing the conversion range especially for the more difficult pentose sugars such as xylose. Protein engineers

are taking a synthetic biology approach with recent progress in engineering more stable and effective enzymes such as cellobiohydrolases by researchers at Caltech [108] and completely novel metabolic pathways by the Berkeley company, Amyris [109].

Transgenic Crops, Next Generation. Table 2 Examples of plant components with suggested functionality^a

Class/components	Source ^b	Potential Health Benefits
Carotenoids		
Alpha-carotene	Carrots	Neutralizes free radicals that may cause damage to cells
Beta-carotene	Various fruits, vegetables	Neutralizes free radicals
Lutein	Green vegetables	Contributes to maintenance of healthy vision
Lycopene	Tomatoes and tomato products (ketchup, sauces)	May reduce risk of prostate cancer
Zeaxanthin	Eggs, citrus, maize	Contributes to the maintenance of healthy vision
Dietary fiber		
Insoluble fiber	Wheat bran	May reduce risk of breast and/or colon cancer
Beta glucan ^c	Oats	May reduce risk of cardiovascular disease (CVD)
Soluble fiber ^c	Psyllium	May reduce risk of CVD
Whole grains ^c	Cereal grains	May reduce risk of CVD
Collagen Hydrolysate	Gelatin	May help improve some symptoms associated with osteoarthritis
Fatty acids		
Omega-3 fatty acids – DHA/EPA	Tuna; fish and marine oils	May reduce risk of CVD and improve mental, visual functions
Conjugated linoleic acid (CLA)	Cheese, meat products	May improve body composition, may decrease risk of certain cancers
Gamma linolenic acid	Borage, evening primrose	May reduce inflammation risk of cancer, CVD disease and improve body composition
Flavonoids		
Anthocyanidins: cyanidin	Berries	Neutralize free radicals, may reduce risk of cancer
Hydroxycinnamates	Wheat	Antioxidant-like activities, may reduce risk of degenerative diseases
Flavanols: catechins, tannins	Tea (green, catechins), (black, tannins)	Neutralize free radicals, may reduce risk of cancer
Flavanones	Citrus	Neutralize free radicals, may reduce risk of cancer
Flavones: quercetin	Fruits/vegetables	Neutralize free radicals, may reduce risk of cancer
Glucosinolates, indoles, isothiocyanates		
Sulphoraphane	Cruciferous vegetables (broccoli, kale), horseradish	Neutralizes free radicals, may reduce risk of cancer
Phenolics		
Stilbenes – resveratrol	Grapes	May reduce risk of degenerative diseases; heart disease; cancer. May have longevity effect
Caffeic acid, ferulic acid	Fruits, vegetables, citrus	Antioxidant-like activities; may reduce risk of degenerative diseases; heart disease, eye disease
Epicatechin	Cacao	Antioxidant-like activities; may reduce risk of degenerative diseases; heart disease

Transgenic Crops, Next Generation. Table 2 (Continued)

Class/components	Source ^b	Potential Health Benefits
Plant stanols/sterols		
Stanol/sterol ester ^c	Maize, soy, wheat, wood oils	May reduce risk of coronary heart disease (CHD) by lowering blood cholesterol levels
Prebiotic/probiotics		
Fructans, inulins, fructo-oligosaccharides (FOS)	Jerusalem artichokes, shallots, onion powder	May improve gastrointestinal health
<i>Lactobacillus</i>	Yogurt, other dairy	May improve gastrointestinal health
Saponins	Soybeans, soy foods, soy protein-containing foods	May lower LDL cholesterol; contains anticancer enzymes
Soybean protein	Soybeans and soy-based foods	25 g/day may reduce risk of heart disease.
Phytoestrogens		
Isoflavones – daidzein, genistein	Soybeans and soy-based foods	May reduce menopause symptoms, such as hot flashes, reduce osteoporosis, CVD
Lignans	Flax, rye, vegetables	May protect against heart disease and some cancers; may lower LDL cholesterol, total cholesterol, and triglycerides
Sulfides/thiols		
Diallyl sulfide	Onions, garlic, olives, leeks, scallions	May lower LDL cholesterol, helps to maintain healthy immune system
Allyl methyl trisulfide, dithiolthiones	Cruciferous vegetables	May lower LDL cholesterol, helps to maintain healthy immune system
Tannins		
Proanthocyanidins	Cranberries, cranberry products, cocoa, chocolate, black tea	May improve urinary tract health May reduce risk of CVD, and high blood pressure

^aExamples are not an all-inclusive list

^bUS Food and Drug Administration approved health claim established for component. Modified from Newell-McGloughlin [29] (25)

Biodiesel is defined by the National Biodiesel Board as a mono-alkyl ester. It is basically vegetable oil or animal fat-based diesel fuel consisting of long-chain alkyl (methyl, propyl, or ethyl) esters. Biodiesel is typically made via a trans-esterification process reacting lipids (vegetable oil, animal fat) with an alcohol.

Biodiesel per se can be used by standard diesel engines and is therefore qualitatively distinct from the vegetable, animal, and other waste oils used to fuel-converted diesel engines. Biodiesel can be used alone, or blended with petrodiesel. Biodiesel has better lubricating properties and much higher cetane ratings than today's lower sulfur diesel fuels but is still not economical as an alternate stand-alone fuel. Although still

carbon-based, it is suggested by some that in terms of biofuels the algal biodiesel approach is much more sustainable than either the cellulosic or other land-based sources with DOE claiming that algae fuel yields have not yet been accurately determined, but DOE is reported as saying that algae yield 30 times more energy per acre than land crops such as soybeans [110].

From a biotechnology perspective the main focus for expanding interest in this area is increasing lipid production and modifying lipid composition for optimum performance. Work is being done to modify algae for increased production of desirable medium chain fatty acids (MCFA) which abrogates the requirement for cracking and isomerizing of long chain fatty acids (LCFAs).

The advantages of MCFAs over LCFAs are high energy density, low fuel viscosity, low flash point, and low freezing point. The real issue with algal production systems is the scale-up step for commercial-level production where contamination, dewetting, and lipid isolation are still economically prohibitive. In February 2010, the Defense Advanced Research Projects Agency announced that the US military was about to begin large-scale production of jet fuel from algal pond isolates [111].

Barriers to Introduction

Most of the crops approved to date do appear to support the notion that the deregulation process is prohibitive for any but well-financed companies whose focus is primarily on the large commodity crops as just discussed. Worldwide there is clear asymmetry and lack of consensus in regulatory systems [112]. This discourages research on anything but the most mundane of crops and traits and is a real disincentive to creative research. For all intent and purposes there is just one trait from a public institution that has successfully traversed the regulatory minefields and been translated into a commercially viable commodity and that is the viral coat protein protection system initially developed for the papaya ringspot virus pandemic in Hawaii. This crop, papaya, is a major tropical fruit crop in the Asian region. However, production in many Asian countries is set back by the prevalence of the PRSV disease as well as post-harvest losses. The PRSV-resistant papaya, based on RNAi suppression of the coat protein expression, literally saved the \$17 M economy in Hawaii and is of significant importance in Taiwan and other SE Asian countries. Coat protein-based resistance is a demonstration of what is known as post-transcriptional gene silencing (PTGS). RNA interference (RNAi) in animals and basal eukaryotes, quelling in fungi, and PTGS in plants are examples of a broad family of phenomena collectively called RNA silencing. This system has now been applied to many species. A 5-year effort to combat plum pox virus disease through PTGS resistance paid off. In 1990, USDA/Agricultural Research Service (ARS) scientists began their efforts with a papaya ringspot virus coat protein gene obtained from Dennis Gonsalves [113]. This gene shows 70% homology to the plum pox gene

and has been used to control other viruses similarly related to papaya ringspot. However, irrespective of the mechanism, it is important that resistance based on a single gene is managed well and alternate control mechanisms are introduced to reduce pressure on the development of viral resistance. Other approaches include expression of the RNA replicating enzymes of the virus, expression of satellite RNA, replicating RNA molecules that are molecular parasites of the virus, or the use of protease inhibitors to interfere with processing of the viral proteins.

While translation of biotech research into value-added products for producers and consumers is a challenge in the USA, it is exponentially more difficult in LDCs [112]. A problem facing Africa in particular is the lack of a dynamic private sector to take technologies to the farmer. It has also been estimated that regulatory costs might exceed the costs of research and experimentation needed to develop a given GM crop, which is a major problem in releasing such crops to the market. A way to reduce the costs of generating food and environmental safety data is to develop regional “centers of excellence” with complementary facilities where food safety testing can be done reliably and regulatory costs could be reduced. The economic gains from using genetically modified crop technology in sub-Saharan Africa (SSA) are potentially large according to the World Bank Group [114]. The results suggest that the welfare gains are potentially very large, especially from golden rice and nutritionally enhanced GM wheat, and that those benefits are diminished only slightly by the presence of the de facto European Union’s current ban on imports of GM foods.

The authors used the global economy-wide computable general equilibrium model known as GTAP. They specifically noted that if SSA countries impose bans on GM crop imports in deference to EU market demand for non-GM products, the domestic consumer loss net of that protectionism boost to SSA farmers would be more than the small gain derived from greater market access to the EU.

Problems cited for the slow passage of GM crops from experimental, to trial, to commercial stage especially in LDCs include the lack of capacity to negotiate licenses to use genes and research techniques patented by others, especially for crops with export potential [28]. In addition, there are difficulties in meeting

regulatory requirements and a lack of effective public commercialization modalities and working extension networks. Biosafety and IPR regulations still have to be enforced in many countries for an effective and safe use of genetically engineered crops, especially if their production is meant for the export market.

Scientific, civic, and religious opinion leaders from all over the world have expressed support for the value of this technology. Florence Wambugu (CEO, Africa Harvest Biotech Foundation International, Kenya) states that the great potential of biotechnology to increase agriculture in Africa lies in its “packaged technology in the seed,” which ensures technology benefits without changing local cultural practices [115]. For example, over 120 million children worldwide are deficient in vitamin A. In the late 1990s, Potrykus [108] group engineered rice to accumulate provitamin A (b-carotene). Incorporation of this trait into rice cultivars and widespread distribution of this “packaged technology in the seed” could prevent one to two million deaths each year. She observes that in the past, many foreign donors funded high-input projects, which have not been sustainable because they have failed to address social and economic issues such as changes in cultural practice [115]. In concurrence with this, Ismail Serageldin, former Chairman of the CGIAR (Consultative Group on International Agricultural Research) noted that, a priori, biotechnology could contribute to food security by helping to promote sustainable agriculture centered on smallholder farmers in developing countries [104].

US Consumer attitudes also tend to be relatively positive on the whole about agricultural biotechnology. In a 2010 IFIC survey, consumers were determined to be largely familiar with the term “biotechnology” [116]: More than two thirds of consumers (69%) have read or heard at least “a little” about the concept. Half (51%) of consumers say they are favorable toward farmers using biotechnology to grow more crops that would help meet food demand. In addition, significantly more consumers this year (28% vs. 23% in 2008) believe foods produced through biotechnology are available in the supermarket today, although this figure is still relatively low. Certain benefits of biotechnology are found to resonate better with consumers than others. These tend to be consumer-facing qualities such as improved health or better taste. For example,

the majority of consumers say they are somewhat or very likely to purchase foods produced through biotechnology to provide more healthful fats like Omega-3 s (76%), to avoid trans fat (74%), and to make foods taste better/fresher (67%). This is consistent from 2008. Additionally, more than three quarters (77%) of consumers say they would be likely to purchase foods produced through biotechnology for their ability to reduce pesticide use, and 73% of consumers said they would be likely to purchase bread, crackers, cookies, cereal, or pasta made with flour from wheat that had been modified by biotechnology to use less land, water, and/or pesticides. Of the 18% who would like to see additional information on the FDA label, only 3% mentioned anything about biotechnology.

But what of the context in which these crops are grown? Can all cropping systems co-exist in harmony? According to Brookes and Barfoot, [8, 117] it is important to determine the relative importance of different crop production systems based on planted area, production, and economic value to the region in question. The issue is what, if any, are the economic consequences of adventitious presence of material from one crop system within another based on the notion that farmers should be able to cultivate freely the crops of their choice using whichever production system works best in any given context (GM, conventional, or organic). It is never a food or environmental safety issue but rather a production and marketing matter. The heart of the issue is assessing the likelihood of adventitious presence of material from one production system affecting another and the potential impacts. This requires consistency when dealing with adventitious presence of any unwanted material including, but most definitely not limited to, biotech-derived material. Adventitious presence is simply the unintended incidence of something other than the desired crop such as small quantities of weed seeds, seeds from other crops, dirt, insects, or foreign material (e.g., stones). It is unrealistic to expect 100% purity for any crops, or products derived therefrom, so thresholds that are consistent across all materials should be set and should not discriminate (e.g., thresholds for adventitious presence of biotech material should be the same as applied to thresholds for other unwanted material and vice versa). All measures should be proportionate, non-discriminatory, and science-based.

The issue of economic liability provisions that compensate growers for adventitious presence of biotech material is often raised [117]. Historically, worldwide the market has adequately addressed economic liability issues relating to the adventitious presence of unwanted material in any agricultural crop. For example, for certified seed the onus is on the producers, who require isolation from undesired pollination for the purity of their product, to insure such purity; this is not their neighbor's problem. By extension the onus is on growers of any specialty crops to take action to protect the purity of their crops since these are self-imposed standards for and by that market. Growers who have themselves chosen a more stringent standard than that established in EU legislation should not expect their neighbors to bear the special management costs of meeting that self-imposed standard; to do so would reverse fundamental freedoms of economic activity and would establish a dangerous precedent. To allow specialty operators to formulate unrealistic standards for GM in their own produce would impose impossibly high standards on neighbors and would effectively impose a ban on the choice of other producers. Such growers usually are rewarded by higher prices and niche markets for taking such actions. Their neighbors enjoy no such advantage.

Existing legislation in North America and the EU is more than adequate to protect all grower and consumer interests but if new regulations were considered to address economic liability provisions for any negative economic consequences of adventitious presence of unwanted material, the same principle should apply to all farmers regardless of their chosen production methods. On equity grounds, biotech growers should have equal access to compensation for adventitious presence of material from conventional or organic crops (such as fungal contamination) as conventional and organic producers have from biotech growers. No one sector should be able to unfairly prohibit another – access and choice work both ways. All co-existence measures should be based on legal, practical, and scientific realities and not on commercial or niche marketing objectives. Where unintended presence has occurred on a number of occasions to date such as the presence of minute levels of Bayer Crop Sciences–regulated material LLRICE 604 found in Clearfield 131 (CL131) rice seed in 2007 and Mycogen's event "32" in

maize in 2008, the agencies cooperated and determined that these events did not prove any risk as they carried similar constructs to those already having achieved non-regulated status.

According to Brookes and Barfoot, [8, 117] biotech crops co-exist successfully with conventional and organic crops in North America (where, as noted, biotech crops account for the majority of acreage of important arable crops like soybeans, cotton, and maize) Spain, and more recently the Czech Republic. The market has developed practical, proportionate, and workable coexistence measures without new regulations or indeed any government intervention. Where isolated instances of adventitious presence of biotech material have been found in conventional or organic crops these have usually been caused by inadequate implementation of good coexistence practices (e.g., inefficient segregation of crops in storage and transport, nonuse of tested, certified seed). Under civil liability (i.e., tort damages) and for intellectual property infringement (except for the unauthorized StarLink), there have been no lawsuits brought by any parties for adventitious presence. Every case brought by a seed company for infringement has involved a claim that the farmer charged with infringement was an intentional infringer (i.e., adventitious presence was not the issue). And, to date, each of these cases was upheld by the courts. Indeed, all except one notable exception in North America has conceded to this claim. The exception is Percy Schmeiser who famously was found by a number of courts to have infringed Monsanto's glyphosate-tolerant patent by deliberately spraying and subsequently saving seeds from resistant sport canola plants found growing near his property. He initially claimed adventitious contamination but upon losing all the way to the Canadian Supreme court he changed tact to try to take advantage of Canadian patent law which prohibits patenting of higher organisms. While the court upheld this they determined that the construct within the plant was subject to IP protection and so Schmeiser was found to have infringed a patentable article under Canadian law.

Virtually all EU member states have transcribed EU Directive 2001/18 and implement EU regulations on traceability and labeling. Within the EU, provision has been made for a de minimis threshold for unavoidable presence of GMOs but no actual threshold has been set.

Therefore, the default state of the 0.9% on labeling and traceability is the one enforced. In the USA, organic products cannot be (legally) downgraded or the producer decertified by unintentional presence when all required measures and best practices are adhered to and no producer has been so impacted to date [118].

Going forward there are four major stanchions to the furtherance of co-existence and all of them are incumbent on cooperation.

1. **Monitoring:** Verify the models and predictions about cost, isolation standards, and generally to learn how the farming community copes with the requirements for keeping the product streams separated.
2. **Dialog:** Strategy development takes place in a dialog between the scientific and technical community and all relevant stakeholders. (The Czech Republic [119].)
3. **Stewardship:** Stewardship programs should take into account the interests of both GM and non-GM farmers. Existing product stewardship programs for non-GM crops in farming should be a starting point for developing stewardship schemes for GM crops.
4. **Research:** The scientific community should be encouraged to fill the knowledge gaps that have been identified. Projects are needed to validate models and guidelines, including long-term studies. Building up mechanistic, probabilistic, and predictive models of gene flow etc. Methods for restricting gene flow by eliminating the fertility of pollen or seeds (apomixis, cytoplasmic male sterility, plastid transformation, Genetic Use Restriction Technology (GURT), etc.).

The World Trade Organization ruled in 2006 that a 6-year European ban on genetically engineered crops violates international trade. The three-person panel issued its decision ruling in favor of the three countries, USA, Canada, and Argentina, on a large majority of the 25 crops under dispute in the case while issuing mixed rulings on a few crops. The panel also ruled in favor of challenging national bans on specific biotech crops issued by Austria, France, Germany, Greece, Italy, and Luxembourg. The EU had argued that it did not have a moratorium but that it just took more time to weigh

the possible risks to health and the environment posed by genetically engineered foods. It said it needed to take a “precautionary” approach to regulation, which is different from what it called Washington’s “laissez-faire” stance.

The trade organization panel appears not to have challenged Europe’s regulatory process for biotech crops. Rather, it said Europe failed to follow its own procedures, resulting in undue delay of decisions. Interestingly one of the most comprehensive assessments on the technology was conducted by EU scientists. An EU Commission Report [120] that summarized biosafety research of 400 scientific teams from all original 15 EU countries conducted over 15 years stated that research on biotechnology-derived plants and products so far developed and marketed, following usual risk assessment procedures, has not shown any new risks to human health or the environment beyond the usual uncertainties of conventional plant breeding. Indeed, the use of more precise technology and the greater regulatory scrutiny probably make them even safer than conventional plants and foods. If there are unforeseen environmental effects – none have appeared as yet – these should be rapidly detected by existing monitoring systems. This analysis was repeated in a 2008 EU Joint Research Centre (JRC) Report commissioned by Members of the European Parliament (MEPs) conducted by world experts including the European Food Safety Authority (EFSA), the World Health Organization (WHO), and others [120]. Their report concluded that there is no evidence that genetically modified foods have any harmful effects; a declaration signed by over 3,500 scientists including 25 Nobel Laureates reiterates this position [28].

Future Directions

As agriculture must adapt to rapidly changing needs and growing conditions it is important to become more effective at producing more or less with limited resources and only the tools of biotechnology will allow us to bypass physiological and environmental limitations to produce sufficient food, feed, fuels, and fiber on ever diminishing arable land to meet ever increasing demand. The challenges going forward are foremost technical as one strives to modify qualitative as opposed to quantitative traits and intricates metabolic pathways and

networks as opposed to single genes; the scientific hurdles to achieve these aims are not trivial. However, with the tools now coming online in the fields of genomics, proteomics, metabolomics, and bioinformatics, there is the potential to make major modifications to introgress desirable traits. For example, tools such as next generation sequencing, RNA interference (RNAi), transcription factors (Tfs), transcription activator-like effector nucleases (TALENs), mini-chromosomes, combinatorial transformation, epigenetic modification, network engineering, and systems biology will allow us to apply both reductive and holistic approaches to identify, modify, introgress, and subsequently and simultaneously study the expression and interaction of transgenes on tens of thousands of endogenous genes in elite germ-plasm backgrounds. With these newly evolving tools, one is beginning to dissect the global effects of metabolic engineering on metabolites, enzyme activities, and fluxes. With rapidly emerging technologies, the increase in our understanding of and ability to manipulate plant metabolism during the coming decades should place plant researchers in the position of being able to modify crop traits to respond to the diversity of needs from minimizing environmental impact to optimizing productivity and quality output.

Non-technical limitations include intellectual property restrictions which may limit translation of public research if not managed judiciously; secondly, liability concerns over abuse or misuse of constructs; thirdly prohibitive and asymmetric biosafety regimes and finally public acceptance. The latter two in many ways are the most insidious of limitations as they have little basis in rational process and thus are difficult to redress effectively – the last in particular is often predicated on how much of the former is perceived to be of concern, and how positions are presented by the opposing factions. It is often easier to appeal to emotion and self-fear than it is to present reasoned and judicious scientific rationale for basing risk analysis. Indeed the actual commercialization of biotech products may have little to do with technical limitations and more to do with these external constraints primarily the process of regulatory approval. The flagship of improved nutritional varieties, namely, beta carotene-enhanced rice commonly referred to as golden rice, despite being under consideration since the late 1990s and subject to a barrage of risk assessments is unlikely to be approved until 2012 at the earliest. Ingo

Potrykus, the developer, says an unreasonable amount of testing has been required without scientific justification. In a recent *Nature* article [121], he lays the blame largely on the regulatory process which he considers excessive observing that unjustified and impractical legal requirements are preventing genetically engineered crops from saving millions from starvation and malnutrition.

In the final analysis, resources are finite and true sustainability can come only from an enlightened philosophy that promotes the development of resource-enhancing technologies. Antithetically, those who claim to be the stalwarts of sustainability are, on occasion, the very ones who oppose the development and application of those tools that can help to insure such sustainability. The only sure way to insure food security and protect the planet's resources is not to settle into the complacency of maintaining the status quo but to engage in continual, constructive change based on scientific knowledge. Thus, if one is to be accountable to posterity it is not just the choice but one's duty to promote and apply responsible science and technology in all endeavors.

Bibliography

1. World Bank (2007) Agriculture for development, overview. World Bank, Washington, DC
2. Lavelle M, Garber K (2008) Eight ways to fix the global food crisis US news and world reports. Posted 9 May 2008. <http://www.climos.com/news/articles/8wayssto.htm>
3. F. W. F. a. A. O. o. t. U. N. W. H. (2004) The state of food and agriculture 2003-2004, agricultural biotechnology, meeting the needs of the poor? vol 35. Food and Agriculture Organization of the United Nations, Rome, pp 1-209
4. Economic and Social Council: Commission on Science and Technology for Development (2004) Resolution 2004/68. Science and Technology for Development. Council, E. a. S
5. Agbioworld (2000) Seven scientific academies support GM crops reproduced from The Times (London), 12 July 2000. <http://www.agbioworld.org/biotech-info/topics/dev-world/academies.html>
6. James C (2011) Global status of commercialized biotech/GM Crops: 2010. ISAAA briefs 42-2010. ISAAA, Ithaca. <http://www.isaaa.org/resources/publications/briefs/42/executivesummary/default.asp>
7. Ghadar F, Spindler H (2005) Industrial management the growth of biotechnology (Global tectonics: Part 1). <http://www.entrepreneur.com/tradejournals/article/135568229.html>
8. Brooks G, Barfoot P (2008) Co-existence of GM and Non-GM arable crops: the Non-GM and organic context in the EU. PG Economics, Dorchester

9. Brooks G, Barfoot P (2005) The global economic and environmental impact – The first nine years 1996–2004. *AgBioForum* 8:187–196
10. Elmegaard N, Pedersen MB (2001) Flora and fauna in Roundup tolerant fodder beet fields. National Environment Research Institute, Technical Report No. 349, p 40
11. Trewavas A (2001) Urban myths of organic farming. *Nature* 410:409–410
12. Folcher L, Delos M, Marengue E, Jarry M, Weissenberger A, Eychenne N, Regnault-Roger C (2010) Lower mycotoxin levels in Bt maize grain. *Agron Sustain Dev* 30(4):711–719
13. Green RE, Cornell SJ, Scharlemann JP, Balmford A (2005) Farming and the fate of wild nature. *Science* 307:550–555
14. Huang J, Rozelle S, Pray C, Wang Q (2002) Plant biotechnology in China. *Science* 295:674–676
15. Choudhary B, Gaur K (2010) BT cotton in India: a country profile. ISAAA, Ithaca
16. Choudhary B, Gaur K (2008) The development and regulation of BT brinjal in India (Eggplant/Aubergine). ISAAA Brief No. 38, Ithaca
17. Bricknell T (2010) India delays BT brinjal decision. *Fruitnet.com*
18. Monsanto (2010) Insect protected roundup ready 2 yield soybeans achieve important milestone toward commercialization in Brazil
19. Stern N (2006) Review on the economics of climate change. HM Treasury, London
20. Svoboda E (2008) The future of farming is in nitrogen efficiency fast company. Posted 1 Oct 2008. <http://www.fastcompany.com/magazine/129/seed-money.html>
21. Sottosanto JB, Saranga Y, Blumwald E (2007) Impact of AtNHX1, a vacuolar Na⁺/H⁺ antiporter, upon gene expression during short- and long-term salt stress in *Arabidopsis thaliana*. *BMC Plant Biol* 7:18
22. Agarwal PK, Agarwal P, Reddy MK, Sopory SK (2006) Role of DREB transcription factors in abiotic and biotic stress tolerance in plants. *Plant Cell Rep* 25:1263–1274
23. Deng X, Phillips J, Brautigam A, Engstrom P, Johannesson H, Ouwerkerk PB, Ruberti I, Salinas J, Vera P, Iannaccone R, Meijer AH, Bartels D (2006) A homeodomain leucine zipper gene from *Craterostigma plantagineum* regulates abscisic acid responsive gene expression and physiological responses. *Plant Mol Biol* 61:469–489
24. Rivero RM, Shulaev V, Blumwald E (2009) Cytokinin-dependent photorespiration and the protection of photosynthesis during water deficit. *Plant Physiol* 150:1530–1540
25. Foundation AAT (2007) Combining breeding and biotechnology to develop water efficient maize for Africa (WEMA)
26. Jung KH, Seo YS, Walia H, Cao P, Fukao T, Canlas PE, Amonpant F, Bailey-Serres J, Ronald PC (2010) The submergence tolerance regulator Sub1A mediates stress-responsive expression of AP2/ERF transcription factors. *Plant Physiol* 152:1674–1692
27. Lopez-Bucio J, Nieto-Jacobo MF, Ramirez-Rodriguez VV, Herrera-Estrella L (2000) Organic acid metabolism in plants: from adaptive physiology to transgenic varieties for cultivation in extreme soils. *Plant Sci* 160:1–13
28. Newell-McGloughlin M (2006) “Functional foods” and biopharmaceuticals: the next generation of the gm revolution. American Enterprise Institute, Washington, DC, pp 163–178
29. Newell-McGloughlin M (2008) Nutritionally improved agricultural crops. *Plant Physiol* 147:939–953
30. Luciani G, Wofford DS (2005) Over-expression of a soybean vegetative storage protein gene in Bahiagrass (*Paspalumnotatum* var. Flugge). http://www.ufgi.ufl.edu/Symposium/FG2005program_rev.pdf
31. Roesler K et al (1997) Targeting of the *Arabidopsis* homomeric acetyl-coenzymeA carboxylase to plastids of rapeseeds. *Plant Physiol* 113:75–81
32. Cromwell GL et al (1967) Nutritional value of opaque-2 corn for swine. *J Anim Sci* 26:1325–1331
33. Yang SH et al (2002) Expression of a synthetic porcine alpha-lactalbumin gene in the kernels of transgenic maize. *Transgenic Res* 11:11–20
34. O’Quinn PR et al (2000) Nutritional value of a genetically improved highlysine, high-oil corn for young pigs. *J Anim Sci* 78:2144–2149
35. Young TE et al (2004) Senescence-induced expression of cytokinin reverses pistil abortion during maize flower development. *Plant J* 38:910–922
36. Chakraborty S et al (2000) Increased nutritive value of transgenic potato by expressing a nonallergenic seed albumin gene from *Amaranthushypochondriacus*. *Proc Natl Acad Sci USA* 97:3724–3729
37. Li LLS et al (2001) Increase of sulphur-containing amino acids in transgenicpotato with 10 kuzein gene from maize. *Chin Sci Bull* 46:482–484
38. Yu J, Ao G (1997) Expression of 10 kDa sulfur-rich prolamin gene of rice in potato. *Acta Bot Sin* 39:329–334
39. Atanassov ABA et al (2004) To reach the poor – results from the ISNAR-IFPRI Next Harvest study on genetically modified crops, public research, and policy implications. EPTD discussion paper – Environment and Production Technology Division, International Food Policy Research Institute, Washington, DC
40. Katsube T et al (1999) Accumulation of soybean glycinin and its assembly with the glutelins in rice. *Plant Physiol* 120:1063–1074
41. Rapp W (2002) Development of soybeans with improved amino acid composition. In: 93 rd AOCS annual meeting and expo. American Oil Chemists’ Society Press, Montreal, pp 79–86
42. Dinkins RDRM et al (2001) Increased sulfur amino acids in soybean plant overexpressing the maize 15 kD zein protein. *In Vitro Cell Dev Biol Plant* 37:742–747
43. Prakash CS, Jaynes J (2000) Increasing the protein content in sweet potato using a synthetic storage protein gene. *Abstr Papers Am Chem Soc* 219:U36
44. Uauy C et al (2006) Gene regulating senescence improves grain protein, zinc, and iron content in wheat. *Science* 314:1298–1301

45. Falco SC et al (1995) Transgenic canola and soybean seeds with increased lysine. *Biotechnology* 13:577–582, New York
46. White CLTL et al (2001) Increased efficiency of wool growth and live weight gain in Merino sheep fed transgenic lupin seed containing sunflower albumin. *J. Sci. Food Agric* 81:147–154
47. Lai JS (2002) Increasing maize seed methionine by mRNA stability. *Plant J* 30:395–402
48. Zeh M et al (2001) Antisense inhibition of threonine synthase leads to high methionine content in transgenic potato plants. *Plant Physiol* 127:792–802
49. Galili G et al (2002) Genetic, molecular, and genomic approaches to improve the value of plant foods and feeds. *Crit Rev Plant Sci* 21:167–204
50. Del Vecchio A (1996) High-laurate canola. How Calgene's program began, where it's headed. *INFORM Int News Fats Oils Relat Mater* 7:230–243
51. Froman B (2002) Genetic modification of oils for improved health benefits: production of long chain omega-3 fatty acids in plants. *Abstr Papers Am Chem Soc* 223:U35
52. James MJ et al (2003) Metabolism of stearidonic acid in human subjects: comparison with the metabolism of other n-3 fatty acids. *Am J Clin Nutr* 77:1140–1145
53. Ursin VM (2003) Modification of plant lipids for human health: development of functional land-based omega-3 fatty acids. *J Nutr* 133:4271–4274
54. Chapman KDA-BS et al (2001) Transgenic cotton plants with increased seed oleic acid content. *J Am Oil Chem Soc* 78:941–947
55. Liu Q et al (2002) High-oleic and high-stearic cottonseed oils: nutritionally improved cooking oils developed using gene silencing. *J Am Coll Nutr* 21(3 suppl):205S–211S
56. Abbadi ADF et al (2004) Biosynthesis of very-long-chain polyunsaturated fatty acids in transgenic oilseeds: constraints on their accumulation. *Plant Cell* 16:2734–2748
57. Parveez G (2003) Novel products from transgenic oil palm. *AgBiotechNet* 5:1–8
58. Jalani BSCS et al (1997) Improvement of palm oil through breeding and biotechnology. *J Am Oil Chem Soc* 74:1451–1455
59. Anai T et al (2003) Improvement of rice (*Oryza sativa* L.) seed oil quality through introduction of a soybean microsomal omega-3 fatty acid desaturase gene. *Plant Cell Rep* 21:988–992
60. Kinney AJ (1998) Designer oils: the high oleic acid soybean. In: Roller SHS (ed) *Genetic modification in the food industry: a strategy for food quality improvement*. Blackie Academic and Professional, Oxford
61. Reddy AS, Thomas TL (1996) Expression of a cyanobacterial delta 6-desaturase gene results in gamma-linolenic acid production in transgenic plants. *Nat Biotechnol* 14:639–642
62. Arcadia Biosciences (2008) Arcadia biosciences and bioriginal food and science corp. Enter strategic alliance to market high GLA safflower oil. *Business Wire*. http://findarticles.com/p/articles/mi_m0EIN/is_2008_Feb_22/ai_n24320185/
63. Smeeckens S (1997) Engineering plant metabolism. *Trends Plant Sci* 2:286–287
64. Sprenger N et al (1997) Fructan synthesis in transgenic tobacco and chicory plants expressing barley sucrose: fructan 6-fructosyltransferase. *FEBS Lett* 400:355–358
65. Sevenier R et al (1998) High level fructan accumulation in a transgenic sugarbeet. *Nat Biotechnol* 16:843–846
66. Caimi PG et al (1996) Fructan accumulation and sucrose metabolism in transgenic maize endosperm expressing a *Bacillus amyloliquefaciens* SacB gene. *Plant Physiol* 110:355–363
67. Hellwege EM et al (1997) Transgenic potato tubers accumulate high levels of 1-kestose and nystose: functional identification of a sucrose 1-fructosyltransferase of artichoke (*Cynara scolymus*) blossom discs. *Plant J* 12:1057–1065
68. Hartwig E et al (1997) Seed protein and its relationship to soluble sugars in soybeans. *Crop Sci* 37:770–773
69. Hellwege EM et al (2000) Transgenic potato (*Solanum tuberosum*) tubers synthesize the full spectrum of inulin molecules naturally occurring in globe artichoke (*Cynara scolymus*) roots. *Proc Natl Acad Sci USA* 97:8699–8704
70. Chiang CYF et al (2005) Expression of a bi-functional and thermostable amylopullulanase in transgenic rice seeds leads to autohydrolysis and altered composition of starch. *Mol Breed* 15:125–143
71. Schwall GP et al (2000) Production of very-high-amylose potato starch by inhibition of SBE A and B. *Nat Biotechnol* 18:551–554
72. Shintani D, DellaPenna D (1998) Elevating the vitamin E content of plants through metabolic engineering. *Science* 282:2098–2100
73. Rocheford TR et al (2002) Enhancement of vitamin E levels in corn. *J Am Coll Nutr* 21(3 suppl):191S–198S
74. Cahoon EB et al (2003) Metabolic redesign of vitamin E biosynthesis in plants for tocotrienol production and increased antioxidant content. *Nat Biotechnol* 21:1082–1087
75. Chen Z et al (2003) Increasing vitamin C content of plants through enhanced ascorbate recycling. *Proc Natl Acad Sci USA* 100:3525–3530
76. Naqvi S et al (2009) Transgenic multivitamin corn through biofortification of endosperm with three vitamins representing three distinct metabolic pathways. *Proc Natl Acad Sci USA* 106:7762–7767
77. Shewmaker CK et al (1999) Seed-specific overexpression of phytoene synthase: increase in carotenoids and other metabolic effects. *Plant J* 20:401–412
78. Ducreux LJ et al (2005) Metabolic engineering of high carotenoid potato tubers containing enhanced levels of beta-carotene and lutein. *J Exp Bot* 56:81–89
79. Ye X et al (2000) Engineering the provitamin A (beta-carotene) biosynthetic pathway into (carotenoid-free) rice endosperm. *Science* 287:303–305
80. Agius F et al (2003) Engineering increased vitamin C levels in plants by overexpression of a D-galacturonic acid reductase. *Nat Biotechnol* 21:177–181
81. Diaz de la Garza R et al (2004) Folate biofortification in tomatoes by engineering the pteridine branch of folate synthesis. *Proc Natl Acad Sci USA* 101:13720–13725

82. Enfissi EM et al (2005) Metabolic engineering of the mevalonate and nonmevalonate isopentenyl diphosphate-forming pathways for the production of health-promoting isoprenoids in tomato. *Plant Biotechnol J* 3:17–27
83. Fraser PDRS et al (2001) Elevation of carotenoids in tomato by genetic manipulation. *J Sci Food Agric* 81:822–827
84. Rosati C et al (2000) Metabolic engineering of beta-carotene and lycopene content in tomato fruit. *Plant J* 24:413–419
85. Szankowski I et al (2003) Transformation of apple (*Malus domestica* Borkh.) with the stilbene synthase gene from grapevine (*Vitis vinifera* L.) and a PGIP gene from kiwi (*Actinidia deliciosa*). *Plant Cell Rep* 22:141–149
86. Hipskind JD, Paiva NL (2000) Constitutive accumulation of aresveratrol-glucoside in transgenic alfalfa increases resistance to *Phomamedicaginis*. *Mol Plant Microbe Interact* 13:551–562
87. Kobayashi SDC et al (2000) Kiwifruits (*Actinidia deliciosa*) transformed with a *Vitis* stilbene synthase gene produce piceid (resveratrol-glucoside). *Plant Cell Rep* 19:904–910
88. Yu O et al (2000) Production of the isoflavones genistein and daidzein in nonlegume dicot and monocot tissues. *Plant Physiol* 124:781–794
89. Lukaszewicz M et al (2004) Antioxidant capacity manipulation in transgenic potato tuber by changes in phenolic compounds content. *J Agric Food Chem* 52:1526–1533
90. Stark-Lorenzen PNB et al (1997) Transfer of a grapevine stilbene synthase gene to rice (*Oryza sativa* L.). *Plant Cell Rep* 16:668–673
91. Yu O et al (2003) Metabolic engineering to increase isoflavone biosynthesis in soybean seed. *Phytochemistry* 63:753–763
92. Giovino G et al (2005) Antioxidant metabolite profiles in tomato fruit constitutively expressing the grapevine stilbene synthase gene. *Plant Biotechnol J* 3:57–69
93. Niggeweg R et al (2004) Engineering plants with increased levels of the antioxidant chlorogenic acid. *Nat Biotechnol* 22:746–754
94. Muir SR et al (2001) Overexpression of petunia chalcone isomerase in tomato results in fruit containing increased levels of flavonols. *Nat Biotechnol* 19:470–474
95. Gonzali S et al (2009) Purple as a tomato: towards high anthocyanin tomatoes. *Trends Plant Sci* 14:237–241
96. UPI (2002) Wheat inhibits colon cancer. United Press International, Washington, DC
97. Austin-Phillips SBE et al (1999) Production of industrial and animal feed enzymes in transgenic Alfalfa. <http://www.molecularfarming.com/nonmedical.html>
98. Goto F et al (2000) Iron accumulation and enhanced growth in transgenic lettuce plants expressing the iron-binding protein ferritin. *Theor Appl Genet* 100:658–664
99. Lucca P et al (2002) Fighting iron deficiency anemia with iron-rich rice. *J Am Coll Nutr* 21(3 suppl):184S–190S
100. Drakakaki G et al (2005) Endosperm-specific co-expression of recombinant soybean ferritin and *Aspergillus* phytase in maize results in significant increases in the levels of bioavailable iron. *Plant Mol Biol* 59:869–880
101. Han G (2009) Origin agritech announces final approval of world's first genetically modified phytase corn. <http://www.allbusiness.com/science-technology/biology-biotechnology-genetic/13453842-1.html>
102. Denbow DM et al (1998) Soybeans transformed with a fungal phytase gene improve phosphorus availability for broilers. *Poult Sci* 77:878–881
103. Brinch-Pedersen H et al (2000) Generation of transgenic wheat (*Triticum aestivum* L.) for constitutive accumulation of an *Aspergillus* phytase. *Mol Breed* 6:195–206
104. Zhang ZB, Kornegay ET, Radcliffe JS, Wilson JH, Veit HP (2000) Comparison of phytase from genetically engineered *Aspergillus* and canola in weanling pig diets. *J Anim Sci* 78:2868–2878
105. Lucca P, Hurrell R, Potrykus I (2002) Fighting iron deficiency anemia with iron-rich rice. *J Am Coll Nutr* 21:184S
106. Newell-McGloughlin M (2006) A Retrospective prospective perspective on agricultural biotechnology ten years on. *J Commer Biotechnol* 12(1):20–27
107. Programme UNE (2009) Towards sustainable production and use of resources: assessing biofuels, United Nations Environment
108. Heinzelman P, Snow CD, Wu I, Nguyen C, Villalobos A, Govindarajan S, Minshall J, Arnold FH (2009) A family of thermostable fungal cellulases created by structure-guided recombination. *Proc Natl Acad Sci USA* 106:5610–5615
109. Ritch E (2008) Microbes drive new Amyris biodiesel plant. Cleantech Group, San Francisco
110. Hartman E (2008) A promising oil alternative: algae energy. *The Washington Post*
111. Goldenberg S (2010) Algae to solve the Pentagon's jet fuel problem. *The Guardian*, London
112. Newell-McGloughlin M (2002) Ten reasons why biotechnology will be important to the developing world. *AgBioWorld* 2(3 and 4)
113. Gonsalves D (1998) Control of papaya ringspot virus in papaya: a case study. *Annu Rev Phytopathol* 36:415–437
114. Anderson K, Jackson LA (2005) Some implications of GM food technology policies for Sub-Saharan Africa. *J Afr Econ* 14(3):385–410
115. Wambugu F (1999) Why Africa needs agricultural biotech. *Nature* 400:15–16
116. Council IFI (2010) Consumer perceptions of food technology. Cogent Research of Cambridge, Massachusetts for the International Food Information Council, Cambridge
117. Brookes G, Barfoot P (2004) Co-existence of GM and non-GM arable crops: the non-GM and organic context in the EU. PG Economics, Dorchester
118. Kershen D (2006) Legal liability issues in agricultural biotechnology. http://www.nationalaglawcenter.org/assets/articles/kershen_biotech.pdf
119. Křístková M (2008) Principles of co-existence. Department of Agriculture, The Czech Republic. http://www.mzp.cz/www/webdav_biosafety.nsf/biosafety/pdf/coexistence.pdf

120. Council EUJR (2008) Scientific and technical contribution to the development of an overall health strategy in the area of GMOs
121. Potrykus I (2010) Regulation must be revolutionized. *Nature* 466:561
122. Boettiger S, Alvarez S, Getting better technologies to the poor. A public intellectual property resource
123. Potrykus I (1999) Vitamin-A and iron-enriched rices may hold key to combating blindness and malnutrition: a biotechnology advance. *Nat Biotechnol* 17:37
124. Report EC (2001) GMO research in perspective
125. Agbios (2008) GM crop database. Agbios. <http://www.agbios.com/dbase.php?action=ShowForm>
126. Zhao ZYK et al (2002) Nutritionally improved transgenic sorghum. *PlantBiotechnology 2002 and beyond*. In: *Proceedings of the 10th IAPTC&B congress, Orlando, 23–28 June 2002*
127. Glei M et al (2006) Both wheat (*Triticumaestivum*) bran arabinoxylans and gutflora-mediated fermentation products protect human colon cells from genotoxic activities of 4-hydroxynonenal and hydrogen peroxide. *J Agric Food Chem* 54:2088–2095

Transgenic Crops, Risk Assessment and Regulatory Framework in the European Union

YANN DEVOS¹, WENDY CRAIG², JOACHIM SCHIEMANN³

¹GMO Unit, European Food Safety Authority (EFSA), Parma, Italy

²Biosafety Unit, International Centre for Genetic Engineering and Biotechnology (ICGEB), Trieste, Italy

³Julius Kühn Institute (JKI), Federal Research Centre for Cultivated Plants, Institute for Biosafety of Genetically Modified Plants, Quedlinburg, Germany

Article Outline

Glossary

Definition of the Subject

Introduction

Regulatory Oversight of GM Plants and Their Derived Food and Feed Products

The Cartagena Protocol on Biosafety

Process-Based Versus Product-Based Approach

Regulatory Framework for GMOs in the EU

Risk Assessment Principles

Risk Assessment Concepts and Approaches

Future Directions

Acknowledgments

Bibliography

Glossary

Biodiversity Biodiversity is the quantity and variability among living organisms within species (genetic diversity), between species and between ecosystems. Biodiversity is not itself an ecosystem service, but underpins the supply of services. The value placed on biodiversity for its own sake is captured under the cultural ecosystem service called “ethical values” (according to the Economics of Ecosystems and Biodiversity report [1]).

Deliberate release Any intentional introduction into the environment of a genetically modified organism (GMO) or a combination of GMOs for which no specific containment measures are used to limit their contact with, and to provide a high level of safety for, the general population and the environment (according to Directive 2001/18/EC on the deliberate release into the environment of GMOs [2]).

Ecosystem An ecosystem is a dynamic complex of plant, animal, and microorganism communities and their non-living environment interacting as a functional unit. Examples of ecosystems include rainforests, grasslands, urban parks, and cultivated farmlands. Ecosystems can be relatively undisturbed by humans, such as virgin rainforests, or can be modified by human activity (according to the Economics of Ecosystems and Biodiversity report [1]).

Ecosystem services Ecosystem services are the benefits that people obtain from ecosystems. Examples include food, freshwater, timber, climate regulation, protection from natural hazards, erosion control, pharmaceutical ingredients, and recreation (according to the Economics of Ecosystems and Biodiversity report [1]).

Environmental harm Environmental harm can be defined as a measurable adverse change in a natural resource or measurable impairment of a natural resource service which may occur directly or indirectly (according to Directive 2004/35/EC on environmental liability [3]).

Genetically modified/transgenic organisms Organisms, such as plants, animals, and microorganisms (with the exception of human beings), in which the genetic material (DNA) has been altered in such

a way that does not occur naturally by mating and/or natural recombination (according to Directive 2001/18/EC on the deliberate release into the environment of GMOs [2]).

Living modified organism Any living organism that possesses a novel combination of genetic material obtained through the use of modern biotechnology (according to the Cartagena Protocol on Biosafety [4]).

Modern biotechnology The application of (1) *in vitro* nucleic acid techniques, including recombinant deoxyribonucleic acid (DNA) and direct introduction of nucleic acid into cells or organelles, or (2) fusion of cells beyond the taxonomic family, that overcome natural physiological reproductive or recombination barriers and that are not techniques used in traditional breeding and selection (according to the Cartagena Protocol on Biosafety [4]).

Organism Any biological entity capable of replication or of transferring genetic material (according to Directive 2001/18/EC on the deliberate release into the environment of GMOs [2]).

Placing on the market Making available to third parties, whether in return for payment or free of charge (according to Directive 2001/18/EC on the deliberate release into the environment of GMOs [2]).

Risk assessment Process of evaluation of risk, including the identification of scientific uncertainties, of the likelihood and severity of an adverse effect(s) or event(s) occurring to human and animal health or the environment following exposure under defined conditions to a risk source(s). A risk assessment comprises problem formulation (or hazard identification), hazard characterization, exposure characterization, and risk characterization (according to [5]).

Definition of the Subject

This contribution describes the risk assessment principles and the regulatory framework for transgenic (genetically modified (GM)) crops in the European Union (EU).

While the global cropping area of GM crops reached 148 million hectares in 2010, the total area

cultivated with GM crops in the EU was less than 100,000 ha. Most GM crops are thus cultivated outside the EU, but might subsequently be imported and eventually further processed in the EU, mostly for animal feed purposes.

It is globally accepted that agro-food biotechnology could contribute to achieving the objectives (conservation of biological diversity, sustainable use of its components, fair and equitable sharing of the benefits arising out of the utilization of genetic resources) laid down in the Convention on Biological Diversity, if developed and used with adequate safety measures for both the environment and human health. Generally, the safety measures are embedded in process- or product-based regulatory framework. The EU regulatory framework is process-based, precautionary, and includes mandatory labeling and traceability requirements for GM crops and their derived food and feed products. During its development, the EU regulatory system has become increasingly more stringent.

GM crops and their derived food and feed products are generally subject to a risk analysis before they can be commercialized. In the EU, the risk analysis consists of three components: risk assessment, risk management, and risk communication. In risk assessment, potential adverse impacts associated with a specific activity are scientifically characterized on a case-by-case basis, while in risk management, policy alternatives to accept, minimize, or reduce the characterized risks are weighed and, if needed, appropriate prevention and control options are selected. Risk management is functionally and temporally separate from risk assessment in order to reduce any conflict of interest and to protect the scientific integrity of risk assessment. Risk communication is defined as an interactive exchange of information and opinions on risk throughout risk analysis, between risk assessors, risk managers, and other interested parties.

When analyzing potential risks, it is important to bear in mind that the real choice is not between GM crops that are inherently risky and traditionally bred ones that are completely safe. The cultivation of existing crops and those with novel traits (including GM crops) will have both positive and negative consequences. To fully acknowledge the overall outcome of adopting specific crops, and to assess and manage more

effectively the environmental footprint of agriculture as a whole, the conclusion is that broader and more balanced legislative oversight is needed in the EU.

Introduction

The global cropping area of GM crops (including soya bean, maize, cotton, oilseed rape, and sugar beet) has consistently increased each year since they were first commercially cultivated in 1996. While the global cropping area of GM crops reached 148 million hectares in 2010, the total area under GM cultivation in the European Union (EU) was approximately 91,400 ha [6]. Most approved GM crops worldwide are thus currently cultivated outside the EU, but might subsequently be imported and eventually further processed in the EU, mostly for animal feed purposes.

The disparity in adoption rates of GM crops between the EU and the rest of the world is generally attributed both to societal and political opposition toward agro-food biotechnology, and to complex regulatory approval procedures in the EU [7, 8]. In the mid-1990s, the advent of GM crops aroused strong societal concerns [9–12]. Fostered by several highly publicized and successive food safety crises (e.g., bovine spongiform encephalopathy, dioxins, emergence of pathogens such as *Escherichia coli* 0157), public suspicion toward regulatory authorities, scientists and technocratic decision making grew [13]. The media, which was explicitly involved in framing public perception and societal image-building of agro-food biotechnology [14, 15], exacerbated the social amplification of risk [16]. In the late 1990s, increasing societal and political opposition contributed to a *de facto* moratorium on new GM crop market approvals. This was adopted at a meeting of the EU Council of environmental Ministers in June 1999, where five EU Member States indicated that they would oppose any new approvals pending a revision of the legislation [17]. The moratorium did not have a formal status and did not revoke pre-1999 approvals of GM crops or food products, nor did it officially prevent new approvals of GM food products. However, the moratorium was *de facto* effective since a sufficient number of Member States ensured a blocking majority of the legislative process [18]. As a consequence, several GM crop

market applications remained blocked in the EU regulatory system for over a decade.

From 1999 onward, policy makers started to continuously revise the legal conditions under which GM crops and derived food and feed products were to be allowed to be used in the EU, in order to slow down further erosion of public and market confidence (reviewed in [10]). These various legal and institutional reforms, although leading to the upheaval of the *de facto* moratorium in 2004 and a regulatory regime that imposes the most stringent criteria for their approval worldwide (cf., WTO dispute between the US and EU), did not dissipate societal concerns. As Gaskell et al. [19] put it “the new regulatory frame appears to have done little to allay the European public’s anxieties about agro-food biotechnology,” and “the years of controversy have led many people in Europe to believe that anything that has to do with GM food is undesirable”. Member States continue to raise safety objections during the approval process for the placing on the market of GM crops. While the new EU regulatory system should guarantee a harmonized and science-based process, none of the GM plant market applications that were positively evaluated by the EU authority responsible for providing advice on the safety of GMOs, the European Food Safety Authority (EFSA), have attained the necessary qualified majority (neither in favor nor against the approval of GMOs) from the relevant regulatory committees or the Council of Ministers, both for exhibiting substantial abstentions in voting [18, 20]. In most cases, the European Commission has adopted favorable decisions that are not endorsed by a qualified majority of Member States, whereas in other cases, decisions are still pending.

In response to the European Commission approvals for the marketing of GM crops, several Member States invoked national safeguard measures to provisionally restrict or prohibit the use and/or sale of approved GMOs in their territory. Even though EFSA concluded that, in terms of risk to human and animal health and the environment, no new scientific evidence had been presented that would invalidate former risk assessments, the Council of Ministers rejected the proposals of the European Commission to repeal invoked safeguard measures. To reduce the recourse of Member States to safeguard measures and to facilitate the decision making process, the European Commission

proposed in July 2010 to confer to Member States the freedom to allow, restrict, or ban the cultivation of GMOs on part or all of their territory [21, 22]. In addition, the European Commission issued a new Recommendation on coexistence of GM crops with conventional and/or organic crops [23] that replaces the previous recommendation from 2003. Whether these proposals will help to unlock the European legal gridlock is debatable. The European Parliament and Council are expected to discuss the proposals, with a view to legal implementation, in the autumn of 2010 [24]. However, several Member States have already notified the European Commission of their intention to prohibit the cultivation of the Amflora® starch potato, which was approved for cultivation in March 2010. It was the first EU approval for cultivation of a GM crop in the past decade. This approval has garnered considerable public controversy [25] and is illustrative of the lasting skeptical and/or ambivalent attitude of the European society toward agro-food biotechnology.

Regulatory Oversight of GM Plants and Their Derived Food and Feed Products

Heightened global awareness and concern over accelerating environmental degradation during the latter quarter of the twentieth century resulted in a desire by the international community to push the protection of the environment higher up the political agenda. These efforts came to fruition in 1992 when the Convention on Biological Diversity (CBD) came into force. Its objectives include “the conservation of biological diversity, the sustainable use of its components and the fair and equitable sharing of the benefits arising out of the utilization of genetic resources” [26]. During the elaboration of the Convention, negotiators recognized that biotechnology could contribute to achieving these objectives, if developed and used with adequate safety measures for both the environment and human health. Accordingly, procedures were developed to address the safe transfer, handling, and use of any LMO (used interchangeable with GMO in this contribution) resulting from biotechnology that may have an adverse effect on the conservation and sustainable use of biological diversity (Article 19.3, CBD). These procedures formed the Cartagena Protocol on Biosafety (CPB; 4), which came into force in 2003 and has 160 signatory

countries to date (October 2010). Parties lacking a cohesive biosafety policy undertook, or are currently undertaking, a number of initiatives to put a national framework in place in order to comply with the CPB.

This period of heightened political activity in environmental protection has coincided with a concomitant rise in GM crop cultivation. The number of countries opting to grow GM crops has increased steadily from 6 in 1996, the first year of commercialization, to 18 in 2003 and 25 in 2009 [27]. Among the top 10 GM crop-growing countries by area, the USA, Argentina, Canada, Uruguay, and Australia are currently not Parties to the CPB. At the same time, many developing countries that have ratified the CPB are still in the process of elaborating a regulatory framework governing the import or cultivation of GM crops. This has led to the current situation where different strategies and standards have been adopted at the national level, caused by the different infrastructures available in developed and developing countries, and has resulted in much confusion and difficulty in harmonizing environment and trade agreements and regulations.

The Cartagena Protocol on Biosafety

The CPB has been the primary driving force behind many countries establishing national biosafety regulatory systems for GM crops and animals. Attempts have been made under the CPB, as an international legally binding treaty, to set forth the scientific and legal boundaries for those systems, and establish a minimum set of rules and procedures to “ensure an adequate level of protection to avoid or minimize potential adverse effects on the conservation and sustainable use of biological diversity, taking into account human health”. The CPB also sets minimum standards for regulating certain aspects concerning the safe transfer, handling, and use of LMOs [28].

The necessary set of techniques to produce GMOs has been defined not only by the CPB, but also by other relevant international treaties, guidelines, and standards, including the Principles for the Risk Analysis of Foods Derived from Modern Biotechnology [29], and the International Standards for Phytosanitary Measures [30]. The application of modern biotechnology allows the intentional crossing of natural breeding

barriers, the underlying molecular processes of which are qualified as sufficiently “new” so that they and the resulting organisms can be patented as inventions.

In many countries, the terms “genetically modified organism,” “genetically engineered organism,” and “transgenic organism” are widely used, including in domestic legislation, to describe LMOs covered by the CPB [31]. Different countries therefore have biosafety regulations to guide the development of GM crops. Such regulations are key to ensuring the environmental and human safety of GMOs and give the public confidence in GM products [28].

Although the CPB covers all LMOs, it primarily addresses two particular uses of LMOs: [1] those that will be intentionally introduced into the environment and [2] those used for food, feed, or processing (FFP). For LMOs used for other purposes, such as LMOs used in the laboratory, the Protocol leaves any regulation to the discretion of the individual country. The CPB also does not cover products derived from LMOs, such as processed foods that have ingredients that came from LMOs.

To ensure the safe transfer, handling, and use of LMOs, the CPB sets up two separate procedures. The first time that an LMO is to be intentionally introduced into the environment, the CPB sets up an Advanced Informed Agreement (AIA) procedure (Article 7). This procedure requires that an exporter of an LMO provides a notice with detailed information about the LMO to the importing country (Article 8). The importing country then reviews the information, conducts a risk assessment, and decides, based on the risk assessment results, whether to approve or reject the LMO (Articles 10 and 15). In deciding whether to accept the LMO, the importing country can invoke risk management measures to address issues that arise from the risk assessment (Article 16). The importing country also can err on the side of precaution (discussed below) and not approve an LMO if there is insufficient information to adequately assess its particular potential risks (Article 10 [6]).

The second procedure set up by the CPB is for LMOs for FFP (such as maize, soya bean, wheat, or other grains that will be fed to humans or animals). For these LMOs, the AIA procedure is not required (Article 11). Instead, the CPB establishes a simpler system

which reflects the decreased likelihood that these LMOs will affect the biodiversity of the exporting country. Before the LMO can be exported to another country, the safety decision in the exporting country is communicated to other countries through the Biosafety Clearing-House (BCH; <http://bch.cbd.int/>). A country may require prior consent, however, under its domestic regulatory framework, as long as that requirement has been posted on the Biosafety Clearing-House (Article 11).

The CPB also contains numerous other provisions that complement the review procedures for LMOs discussed above and address issues important to a uniform and comprehensive biosafety regulatory process. There are provisions on reviewing decisions for new information (Article 12), simplified procedures for certain LMOs that do not present risks (Article 13) and emergency procedures for unintentional releases of LMOs (Article 17). The CPB also addresses issues such as public awareness and participation (Article 23), and what to do about confidential information (Article 21). Thus, the Protocol attempts to establish a complete and comprehensive set of procedures and legal obligations to assess and manage the potential risks of LMOs on biological diversity, also taking into account risks to human health.

While proponents of modern biotechnology state that no new risks are associated with GMOs, others feel that the new methods of producing organisms might be associated with new risks. Premarket procedures have therefore been established by many regulatory authorities around the world, and are applied to assess how the organisms may behave and evolve in the environment, and how they may interact with other species. The CPB sets forth the information about an LMO that is needed before it is released into the environment or used for FFP. Annexes I and II contain detailed lists on the major categories of information needed to assess the potential risks of an LMO. They provide models which a national biosafety regulatory system can use as standard data requirements. Of course, individual countries may add to the list of required information, depending on particular environment issues within their country or if they choose to address other risk areas (such as food safety or socioeconomic concerns).

The CPB also attempts to explain what a scientific risk assessment of an LMO should entail. Annex III sets forth the objective of the risk assessment, what the risk assessment will be used for, the general principles that the risk assessment must follow, the methodology of the risk assessment, and particular points to consider when assessing the potential risks of an LMO. The Annex provides a clear explanation to interested parties about what is expected in the risk assessment, what will guide the risk assessment, and how it will be used. Therefore, the Protocol's Annexes attempts to provide sufficient information and details so that countries which adopt those provisions will establish harmonized and standardized procedures that will be transparent and understandable.

Precaution and the CPB

Based on the reaffirmation of the precautionary approach contained in Principle 15 of the Rio Declaration on Environment and Development (1992), Articles 10 (para. 6) and 11 (para. 8) of the CPB both state that

- "Lack of scientific certainty due to insufficient relevant scientific information and knowledge regarding the extent of the potential adverse effects of a living modified organism on the conservation and sustainable use of biological diversity in the Party of import, taking also into account risks to human health, shall not prevent that Party from taking a decision, as appropriate, with regard to the import of the living modified organism . . . in order to avoid or minimize such potential adverse effects".

However, ever since the CPB came into force, government regulators and their technical experts, political activists, and GM product developers have debated this inclusion of the precautionary approach. Disagreements have raged over whether such an approach is a useful tool for managing the risks of biotechnology and its products. Resolving these disputes is made difficult by the lack of a formal, established definition, making it unclear exactly what it means in practical terms and what it requires of governments and innovators. Therefore, governments may act at their discretion to restrict or ban products or activities even before

obtaining proof that a harm is imminent, although no obligation to do so seems to be implied in the CPB. This had led opponents of such actions to argue that conversely, such decisions can also jeopardize human health and the environment at large, for example, during the severe food shortage of 2002 in Southern Africa when food aid shipments containing transgenic maize were rejected on the basis of potential harm. Until clarity is forthcoming, this will remain fertile ground for dispute and open for possible abuse.

Process-Based Versus Product-Based Approach

In Europe, as in all Parties of the CPB, a *process-based* regulatory system governs the regulation of GMOs, as the techniques used for their production were considered new and raised specific safety concerns. A GMO is thus mainly characterized by the technique used to produce it, and is defined as an organism in which the genetic material has been altered in a way that does not occur naturally by crossing and/or natural recombination [2]. Directive 2001/18/EC, on the deliberate release of GMOs into the environment, provides a list of techniques in Annexes IA and IB that: (1) result in genetic modification; (2) are not considered to result in genetic modification; or that (3) result in genetic modification, but yield organisms that are excluded from the scope of the Directive [2, 32]. Production techniques falling under the EU GMO definition are:

- Recombinant nucleic acid techniques involving the formation of new combinations of genetic material by the insertion of nucleic acid molecules produced by whatever means outside an organism, into any virus, bacterial plasmid, or other vector system and their incorporation into a host organism in which they do not naturally occur but in which they are capable of continued propagation
- Techniques involving the direct introduction into an organism of heritable material prepared outside the organism including microinjection, macroinjection, and microencapsulation
- Cell fusion (including protoplast fusion) or hybridization techniques where live cells with new combinations of heritable genetic material are formed through the fusion of two or more cells by means of methods that do not occur naturally

In vitro fertilization, natural transformation processes (e.g., conjugation, transduction, transformation), and polyploidy induction are not considered to result in genetic modification and are, therefore, currently excluded from the GMO definition. Techniques of genetic modification yielding organisms that currently are excluded from the GMO definition are mutagenesis and cell fusion of plant cells of organisms which can exchange genetic material through “traditional” production techniques [2].

In the USA and Canada, a *product-based* regulatory approach is followed for the regulation of GMOs [33–35]. Legislation focuses on the risks of products, and not the techniques of production, as genetic modification *per se* is not considered inherently risky. Because the focus is on novel traits or attributes introduced into a plant, rather than the method of production, both plants and their derived food and feed products are regulated under the existing regulatory system.

Whether legal regimes using production techniques instead of the product itself as a trigger for regulatory oversight provide the best framework for an adequate safety assessment of GMOs is at best debatable. On the one hand, it is questionable whether newly developed crop improvement techniques will outgrow the current EU GMO legislation [36, 37]. On the other hand, it is not intuitively obvious why conventionally bred plants and their derived products are not subject to a similar safety assessment as those obtained through genetic modification [38, 39] or, conversely, why GM plants are regulated more strictly than conventionally bred ones in the EU [40, 41]. With techno-scientific advances and innovations, the knowledge about plant genes and their regulation and functions has increased, while new plant production techniques have emerged to induce or select desired plant characteristics, such as RNA interference and oligonucleotide-mediated mutagenesis (see, e.g., [32, 42–44]). These new techniques may challenge the current regulatory definition of a GMO, because it is not always clear whether the products obtained through these techniques would be captured by, or excluded from, the EU GMO definition [36]. The current EU regulation is therefore heavily reliant upon the need to regularly update the list of techniques and their possible uses; otherwise, it runs the risk of quickly becoming obsolete, as the rate of

innovation and advances in the field of biotechnology marches on. In addition, the ability to detect products of such emerging technologies will be severely challenged, and as such, there will need to be a greater emphasis placed on traceability mechanisms for this approach to be practically regulated. In recognition of this, EU Member States and the European Commission are considering recent developments in plant breeding and are discussing if the current EU legalization appropriately covers these new techniques and their application, and whether they should be subject to regulatory requirements. Recently, in an answer to a parliamentary question, the European Commission stated that a specific working group of external experts has been put in place to determine which of the newly developed plant production techniques would result in genetic modification and would thus be captured by, or excluded from, the EU GMO definition (cf., Parliamentary question P-6606/07 2008).

Regulatory Framework for GMOs in the EU

The EU has the most stringent and wide-ranging regulations on GM products and commodities in the world. Their development can be traced back to the aforementioned food safety incidences in the mid-1990s that affected European consumer confidence in government regulatory agencies and agribusiness groups. These concerns have developed to include a negative view of GM products and of the companies that develop and market these products. Furthermore, the ongoing maintenance of trade barriers against agricultural imports in general has resulted in strong political pressure to regulate GM products [45].

The EU regulatory approach is precautionary, process-related, and includes mandatory labeling and traceability requirements for food and feed crops, unprocessed or processed. Only non-food GM products (unseeded), such as textile or other industrial products, are not subject to any requirement.

EU legislation is adopted through a system of interactions between the three main EU institutions: the European Parliament; the Council of the European Union (i.e., representatives of all the EU Member States at the ministerial level); and the European Commission. In most cases, the European

Commission initiates legislative proposals that are decided jointly by the Council and the European Parliament; the most common legislative measures are Regulations (acts that are binding in their entirety and are immediately applicable throughout the EU) and Directives (acts that require the modification or establishment of national measures, generally for harmonization purposes). Legislation pertaining to GMOs in the EU is further elaborated below.

Contained Use and Deliberate Release Directives

In the early 1990s, two European Directives for the use of GMOs were adopted: to ensure the protection of human and animal health and the environment; to guarantee consumers' freedom of choice without misleading consumers/users; and to create an internal market that makes the free movement of GMOs possible within the EU without unequal competition and impediments between and within Member States. Directive 90/219/EEC, which has since been amended by Directive 98/81/EEC, regulated the contained use of GM microorganisms, while Directive 90/220/EEC regulated the deliberate release of GMOs into the environment, covering both the release for research purposes (part B) and for commercial use as or in products (part C). This triad reflects the stepwise process GM crops go through, beginning with experiments under contained use (e.g., laboratory, greenhouse), through experimental release, up to the placing on the market. According to the step-by-step principle, the containment of GMOs can be reduced and the scale of release increased gradually, if assessment of earlier steps indicated that the next step can be taken.

Novel Food Regulation

On 15 May 1997, Regulation (EC) No 258/97 – the so-called *Novel Food Regulation* – removed food products derived from GM plants from the scope of Directive 90/220/EEC on the deliberate release of GMOs into the environment. The new regulation covered risk assessment procedures and the marketing and labeling of all types of novel foods, including those produced by new plant production techniques such as genetic engineering, as well as food without a history of safe use in the EU. Under the Novel Food Regulation, the safety

assessment of GM food was based on the principle of *substantial equivalence* between the GM food and its traditionally cultivated non-GM counterpart. The non-GM counterpart was generally taken to have a history of safe use, allowing it to serve as baseline in the comparison of its chemical composition and phenotypic characteristics to those of GM food [46–48].

According to the *labeling provisions* of the Novel Food Regulation, labeling was not required when a GM raw material had been treated technically in such a way that neither the new DNA, nor the protein could be detected in the end product. Since May 1997, processed oil from GM oilseed rape, maize, and cotton, and processed food and food ingredients derived from GM maize have been notified as being substantially equivalent and thus approved for human consumption under the simplified procedure of the Novel Food Regulation. Moreover, labeling did not apply to food already used for human consumption in the EU prior to the establishment of the Novel Food Regulation. Food already marketed, such as GTS40-3-2 soya bean and Bt176 maize, were not considered as novel. With the adoption of Regulations (EC) No 1813/97 and 1139/98, the labeling of GTS40-3-2 soya bean and Bt176 maize foodstuffs also became compulsory. From that moment on, the label literally had to contain the words “produced from GM soya bean” or “produced from GM maize” when the new protein or transgene was detectable in the end product intended for consumption. A final product needed no label when a GM raw material had been technically treated in such a way that neither the new protein nor the transgene could be detected (e.g., hydrolyzed soya bean proteins, refined oils). With Regulation (EC) No 50/2000, the labeling provisions were extended to additives and flavorings that have been genetically modified or that have been produced from a GMO.

Revised Deliberate Release Directive

On 17 October 2002, Directive 2001/18/EC replaced (the older) Directive 90/220/EEC. With it (1) the precautionary principle was explicitly adopted as a guide; (2) risk assessment criteria were broadened to include direct, indirect, immediate, delayed, and cumulative long-term adverse effects; (3) post-market environmental monitoring (PMEM) became compulsory;

(4) the need for a common methodology for the environmental risk assessment was established; (5) the requirement of reexamination of risk assessment and management conclusions in light of new scientific evidence was strengthened by limiting the duration of market consents to a maximum of ten years; (6) specific considerations related to the use of antibiotic resistance marker genes were introduced; (7) existing labeling provisions applying to GM food were extended to all marketed products containing GMOs; (8) the general concept of traceability at all stages of commercialization was introduced; (9) transparency in the decision making process was increased; (10) consultation of the public became mandatory in the approval procedure; (11) the possibility to consult an ethics committee was confirmed; and (12) the implementation of national cultivation registers was required, recording the locations where GM plants have been grown for experimental purposes.

General Food Law and Establishment of the European Food Safety Authority

Adding to Directive 2001/18/EC, Regulation (EC) No 178/2002 – the so-called *General Food Law* – laid down general principles of food law and procedures in food and feed safety. It defines food and feed and other agricultural inputs at the level of primary production, as well as hazard, risk, risk analysis, risk assessment, risk management, and risk communication. Furthermore, the General Food Law sets food and feed safety requirements in order to determine whether any food or feed may be injurious to human and animal health. With this Regulation, the application of the precautionary principle was further extended to risk analysis of all food and feed in the EU, whether of GM-origin or not.

In response to a multiple wave of food crises that caused considerable public concern in Europe about food safety and the ability of regulatory authorities to fully protect consumers, the *European Food Safety Authority* (EFSA) was created as a European-wide risk assessment body. EFSA is tasked: (1) to provide science-based advice at the request of the European Commission on any matter within its mission, or in the framework of Community legislation; (2) to issue scientific advice on its own initiative; and (3) to issue advice upon request of the European Parliament

or a Member State on matters falling within its mission (Article 29 of the General Food Law). By providing “independent, objective, and transparent” science-based advice, EFSA aims to ensure a high level of consumer protection and to restore and maintain confidence in the EU food supply.

EFSA has ten Scientific Panels addressing food safety issues in the different sectors of food and feed production, and a Scientific Committee. The EFSA Scientific Panel on Genetically Modified Organisms (EFSA GMO Panel) consists of 21 scientific experts from European research institutes, universities, or risk assessment bodies. The EFSA GMO Panel issues: (1) *scientific opinions* on the safety of GMO market approval dossiers and on national safeguard clauses invoked by Member States; and (2) produces *guidelines* for the risk assessment of GMOs upon request of the European Commission in the framework of Directive 2001/18/EC and Regulation (EC) No 1829/2003. These guidelines provide assistance to those preparing and presenting GMO market applications, by describing principles, concepts, data requirements, and issues to be considered in the frame of risk assessment. So far, the following guidances have been issued:

- Guidance document for the risk assessment of GM plants and derived food and feed (issued in 2004; published in 2006; revised in 2008 and 2011)
- Guidance document for the risk assessment of GM microorganisms and their derived products intended for food and feed use (issued in 2006; revised in 2011)
- Guidance document for the renewal of authorizations of existing GMO products (issued in 2006)
- Guidance document for the risk assessment of GM plants containing stacked transformation events (issued in 2007; replaced by the latest revision of the Guidance document for the risk assessment of GM plants and derived food and feed, see above)
- Guidance for the risk assessment of GM plants used for non-food or non-feed purposes (issued in 2009)
- Guidance on the environmental risk assessment of GM plants (issued in 2010)
- Guidance on selection of comparators for the risk assessment of GM plants (issued in 2011)
- Guidance on the PMEM of GM plants (issued in 2011)

Scientific opinions and reports are also issued on specific risk assessment issues regarding food, feed, and environmental safety related to GMOs, such as: the safety of antibiotic resistance marker genes; the use of animal feeding trials for testing of whole GM food and feed; the statistical analysis of results of field trials; the evaluation of GM plants cultivated for non-food/feed purposes; PMEM; statistical considerations for the safety evaluation of GMOs; and the assessment of allergenicity of GM foods. A scientific opinion on the environmental risk assessment of non-target organisms exposed to GM plants was recently adopted, as well as a scientific opinion on the annual PMEM report on the cultivation of maize MON810 in 2009. The EFSA GMO Panel is continuously considering new scientific information and recent developments in the field of GMO risk assessment, and as such, it has undertaken several initiatives to incorporate them into its operational procedures (see [49, 50] for further details).

GM Food and Feed Regulation

Issued on 18 April 2004, Regulation (EC) No 1829/2003 on GM food and feed covers the commercialization and risk assessment of GM food and feed, such as food/feed containing or consisting of, food/feed produced from, and food/feed containing ingredients produced from GMOs, as well as seed or other plant-propagating material. Prior to this date, approvals for human food use were required under the Novel Food Regulation, whereas feed use was assessed under Directive 2001/18/EC and its predecessor. The amended approval procedure is centralized around EFSA and based on a “one door – one key” approach whereby all commercial uses can be covered in the same GM plant market approval dossier. Moreover, it also introduces the need for a GM crop market application to cover both food and feed uses, as it avoids market approval for a single use in case a product is likely to be used for both purposes (e.g., [51]). In response to the heated debates that challenged the “principle of substantial equivalence,” the principle was demoted to a “comparative analysis” in the GM Food and Feed Regulation. The safety assessment of GM food and feed remains based on a comparative analysis in which the non-GM counterpart serves as baseline. However, it requires more evidence of safety than before. Since the principle

of “substantial equivalence” was intensely criticized by various actors, it was no longer interpreted as the endpoint of risk assessment, but rather as the starting point. Whether GM food is subject to further safety assessment depends on the identified similarities and differences between the GM food and its non-GM counterpart: (1) if substantial equivalence is established, the need for further testing is to be investigated on a case-by-case basis; (2) if substantial equivalence is established, except for a single or few specific traits of GM crops, further tests must be done focusing on these traits in order to assess their potential impact on human and animal health; and (3) if neither partial nor total substantial equivalence is established, the wholesomeness of the food product is to be assessed. In the same line of arguments, the simplified authorization procedure was abandoned under the GM Food and Feed Regulation.

Approval Procedures for placing GM crops on the Market

Procedures for approval of GM crops in the EU follow comitology rules that involve different actors such as the European Commission, EFSA, the Regulatory Committee of Member States representatives under Directive 2001/18/EC, the Standing Committee on the Food Chain and Animal Health (SCFCAH) under Regulation (EC) 1829/2003, and the Councils of Ministers. In principle, the delegation of powers to the European Commission along with the supervision of the European Commission’s use of these powers through committees composed of Member States representatives is considered a convenient mechanism. It enables efficient decision making, engenders a close and cooperative working relationship between the European Commission and Member States, and maintains a degree of control over the process by Member States [20]. The voting system is based on the number of votes designated to each of the 27 Member States and is proportional to their percentage of the EU population. A qualified majority should be at least 74% of the total votes (255) and 14 Member States and 62% of the EU population; a blocking minority is at least 26% of the votes [91] or 14 Member States or >38% of the EU population [18].

Approved GM products can move freely throughout all EU Member States in conformity with any

conditions set out in the approval, and enter into the public register of GM food and feed (Community Register: http://ec.europa.eu/food/dyna/gm_register/index_en.cfm). Authorizations are valid throughout the EU, have a maximum duration of ten years, and can be renewed.

Approval Procedure Under Directive 2001/18/EC

Under Directive 2001/18/EC, the approval procedure for placing a GMO on the market involves all Member States, the European Commission, and possibly EFSA (see Fig. 1). The GM crop market application is first submitted to the National Competent Authority of a Member State. Upon receipt of the application, the National Competent Authority issues a risk assessment report that may be favorable or unfavorable regarding market approval of the GMO under consideration. In the event of a favorable opinion, the Member State informs the other Member States via the European Commission. The other Member States and the European Commission examine the risk assessment report and may pose observations and objections.

If there are no objections by other Member States or by the European Commission, the National Competent Authority that carried out the original risk assessment approves the placing on the market of the product, and issues a final approval permitting the marketing of the GMO within the Community.

If objections are raised, the procedure provides for a conciliation phase among the Member States, the European Commission, and the applicant in order to resolve the outstanding questions. In case objections are maintained at the end of the conciliation phase, a Decision is taken at EU level. The European Commission first asks the scientific opinion of the EFSA GMO Panel on the safety of the GMO in question. In the case of EFSA issuing a positive opinion, the European Commission presents a Draft Decision to the Regulatory Committee under Directive 2001/18/EC for an opinion. If the Regulatory Committee gives a favorable opinion by qualified majority, the European Commission adopts the Decision. If not, the Draft Decision is submitted to the Council of Ministers for adoption or rejection by qualified majority. If the Council does not act within three months, the European Commission adopts the draft Decision.

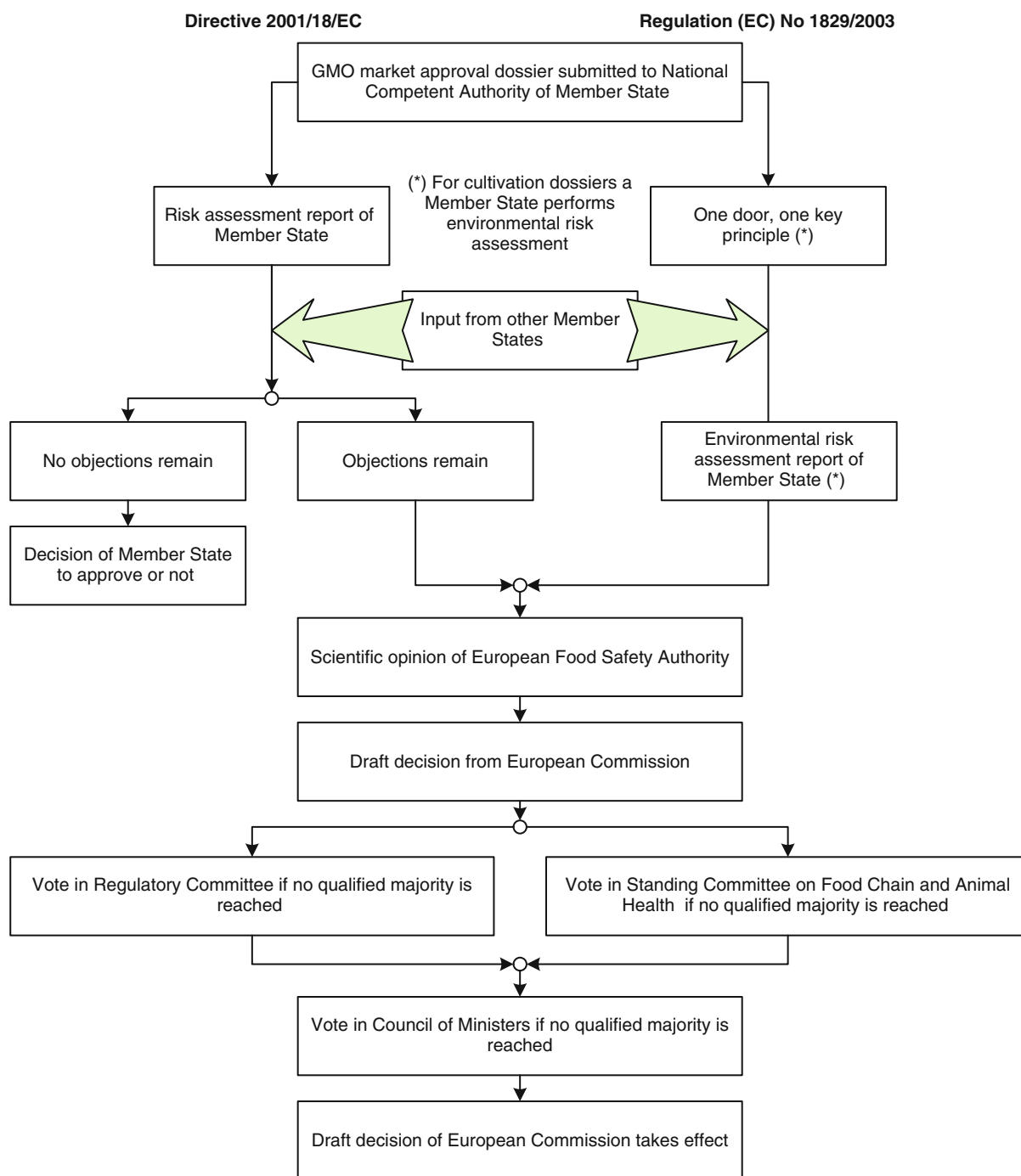
Approval Procedure Under the GM Food and Feed Regulation

Under the GM food and feed Regulation, a GM crop market application is submitted to the National Competent Authority of a Member State, which forwards it to EFSA. If the application covers the use of seeds or other plant-propagating material for cultivation, EFSA asks a National Competent Authority under Directive 2001/18/EC to perform the environmental risk assessment that will be considered by EFSA during its final assessment (see e.g., [52]).

From the receipt of a valid GM crop market application, EFSA endeavors to deliver its scientific opinion within a time limit of six months. This time limit can be extended if EFSA or the Commission's Community Reference Laboratory seeks supplementary information from the applicant. Within three months after receiving EFSA's scientific opinion, the European Commission submits a Draft Decision to the SCFCAH. If SCFCAH gives a favorable opinion by qualified majority, the European Commission adopts the Decision. If not, the Draft Decision is submitted to the Council of Ministers for adoption or rejection by qualified majority. If the Council does not act within three months or does not obtain a qualifying majority for adoption or rejection of the Commission Draft Decision, the European Commission adopts the Decision (see Fig. 1).

Labeling and Traceability Regulation

Regulation (EC) No 1830/2003 on the labeling and traceability of GM food and feed complements, clarifies, and makes operational some of the labeling and traceability objectives of previous legislation. This Regulation extended labeling provisions to feed, seed, bulk products, food that are delivered as such to the final consumers or mass caterers, and to products in which GMO-derived DNA or protein (e.g., refined oils) is no longer detectable. These requirements go further than previously, because the use of genetic modification in itself is now sufficient to justify labeling. The label must be shown in a clearly visible, legible, and indelible manner, and must contain the reference "genetically modified," "produced from genetically modified" or "contains genetically modified." When a GM food or feed is different from its conventional counterpart, the label must also provide information about any characteristic or property that renders it different. When there



Transgenic Crops, Risk Assessment and Regulatory Framework in the European Union. Figure 1

Approval procedure for placing GM crops on the EU market under Directive 2001/18/EC and Regulation (EC) No 1829/2003 (Figure adapted from EFSA (Parma, Italy) and GHK Consulting Ltd (London, UK))

is no conventional counterpart, the label must contain appropriate information about the nature and characteristics of the GM food or feed. Any characteristic or property that gives rise to ethical or even religious concerns also has to be mentioned. Products being produced with the help of GMOs – rather than actually made out of them – do not require labeling. As such, meat, eggs, milk, and dairy products from animals fed GM crops fall outside the remit of labeling provisions. Since substances assisting in food production, carrier substances, or culture media for microorganisms are not considered foods, their labeling is not considered necessary.

Labeling Thresholds

Tolerance thresholds for the unintentional or technically unavoidable presence of approved GM material in non-GM products were established in the EU. A tolerance threshold refers to the maximum admixture level for GMO content under which the comingled product does not have to be labeled as containing a GMO. It is often argued that there is no scientific justification for the established thresholds. Since GM crops and derived food and feed products are declared safe before marketing, thresholds do not relate to safety or health issues. However, thresholds reflect a balance between differently framed societal concerns and requests and technical capabilities. The translation of labeling thresholds in practice still entails enormous technical and scientific challenges.

Regulation (EC) No 49/2000 set the labeling threshold for the adventitious GMO presence in non-GM food at 1% of the food ingredient. The GM Food and Feed Regulation decreased the tolerance threshold to 0.9%, and extended it to feed and products intended for direct processing. There is zero tolerance in the EU concerning unapproved GM plant events, unless they had previously received a favorable scientific risk assessment for marketing from EFSA and a detection method is publicly available. In the latter case, a threshold of 0.5% may be applied transitionally. A threshold has yet to be defined for seeds as discussions have remained at an impasse. Possible thresholds that will be proposed for seeds will be established at levels such that the GMO content of 0.9% can be guaranteed in food, feed, or crops. Proposals made by

the Scientific Committee on Plants in 2001 ranged between 0.3% for cross-pollinating crops, and 0.5% for self-pollinating and vegetatively propagated crops [53]. As no such thresholds have been established to date, any seed lot containing authorized GM seed for cultivation in the EU has to be labeled as containing GM material.

Organic growers principally aim at keeping their products free from any GM material. Regulation (EC) No 1804/1999 on organic production of agricultural products states that the use of GMOs is not compatible with the organic production method. The Regulation, however, foresees a *de minimis* tolerance threshold for the unavoidable presence of GM material in organic products. It was thus anticipated that organic producers would opt for a tolerance threshold ranging between the limit of quantification of a DNA analysis (0.1%) and the tolerance threshold for food and feed products (0.9%). In a press release published on 21 December 2005 (IP/05/1679), the European Commission emphasized that an organic product with an adventitious content of GM material below 0.9% could still be labeled as organic. On 12 June 2007, this point of view was confirmed at a meeting of the EU agriculture ministers, where political agreement was reached on a new Regulation on organic production and labeling (IP/07/807). However, in its recent recommendation on guidelines for the development of national coexistence measures to avoid the unintended presence of GMOs in conventional and organic crops [23], the European Commission allows Member States to aim at levels of unintended GMO presence that are *lower* than the 0.9% labeling threshold in certain cases. It stated that

- “The potential loss of income for organic and some conventional producers (e.g., certain food producers) may be due to the presence of GMO traces at levels lower than 0.9%. In those cases, and in the interest of protecting particular types of production, concerned Member States may define measures that aim at reaching levels of presence of GMOs in other crops lower than 0.9%”.

In this respect, the European Commission referred to Member States that have developed national standards for different types of “GM-free labeling” [23]. Since the organic sector advocates that GM crops are not

compatible with organic farming [54, 55], they will seek to establish the limit of DNA quantification analysis as the basis to determine the tolerance threshold in organic products.

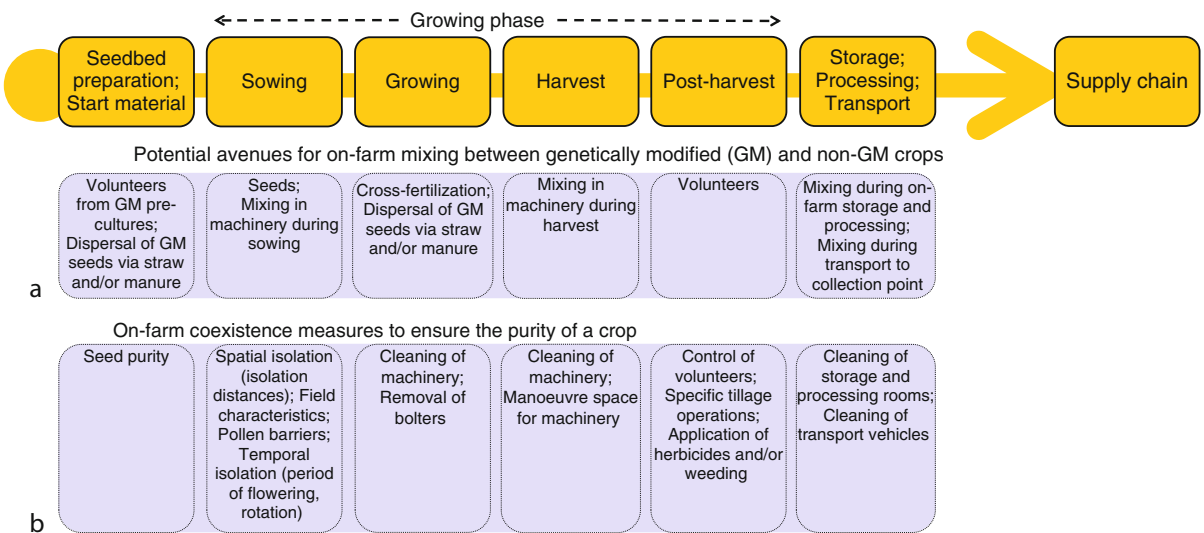
Coexistence

In order to provide a high degree of consumers' choice in the EU, a coexistence policy was adopted to maintain different agricultural production systems. It specifically aimed to enable the side-by-side development of different cropping systems without excluding any agricultural option. In this way, farmers would maintain their ability to make a practical choice between conventional, organic, and GM crops. Since coexistence only applies to approved GM crops that were judged to be safe before their market entry [56], safety issues fall outside the remit of coexistence [57–59].

The European Commission recognized that completely avoiding the unintentional presence of GM material in non-GM products is difficult in the agricultural context [23, 60]. As agriculture is an open system, a certain amount of adventitious mixing is unavoidable. Various sources have been identified that could contribute to on-farm adventitious mixing between GM and non-GM crops (Fig. 2): (1) the use

of impure seed [61, 62]; (2) cross-fertilization due to pollen flow between neighboring fields [63–65]; (3) the occurrence of volunteer plants originating from seeds and/or vegetative plant parts from previous GM crops [66–69]; (4) mixing of plant material in machinery during sowing, harvest, and/or postharvest operations [51]; and – to a lesser extent – (5) cross-fertilization from certain sexually compatible wild relatives and feral plants [70–72]. As completely avoiding admixing is difficult in the agricultural context, tolerance thresholds were established for the unintentional or technically unavoidable presence of approved GM material in non-GM products. If the content of GM material in a non-GM product exceeds the established tolerance threshold of 0.9%, the product has to be labeled as containing GM material, which may affect its market acceptability.

According to Article 43 of the GM Food and Feed Regulation, Member States are empowered to take appropriate measures to avoid the unintentional presence of GM material in other products. There are principally two strategies Member States have established, or are developing, to warrant coexistence of different cropping systems: *ex ante* regulations; and *ex post* liability schemes [23, 73–75]. Regulations are considered *ex ante* if they have to be followed by GM



Transgenic Crops, Risk Assessment and Regulatory Framework in the European Union. Figure 2 (a) Potential avenues for on-farm adventitious mixing between GM and non-GM crops and (b) on-farm coexistence measures to ensure the purity of a crop during the production process (Figure reprinted from [59])

crop adopters while growing GM crops. *Ex ante* regulations prescribe preventive on-farm measures that should ensure that tolerance thresholds are not exceeded in neighboring non-GM agricultural production systems. Contrary to *ex ante* coexistence regulations, *ex post* liability schemes are backward looking: they cover questions of liability and the requirement to redress the incurred economic harm once adventitious mixing in a non-GM product has occurred after the cultivation of GM crops [59].

For decades, seed production regulations have specified statutory segregation measures (so-called identity preservation measures) between seed crops and conventional crop production of the same species to maximize varietal seed purity. Apart from seed production, experience with identity preservation systems is also available from the cultivation of different crop types grown for different uses [76]. Several of the proposed measures to ensure varietal seed and crop purity can be applied within the context of coexistence to limit the adventitious content of GM material in seeds and plant products. These measures include: (1) the use of certified seed; (2) spatially isolating fields of the same crop; (3) implementing pollen barriers around fields; (4) scheduling different sowing and flowering periods; (5) limiting carryover of GM volunteers into the following crop through the extension of cropping intervals; (6) cleaning agricultural machinery and transport vehicles of seed remnants; (7) controlling volunteers and wild relatives; (8) applying effective postharvest tillage operations; (9) retaining records of field history; and (10) the voluntary clustering of fields (Fig. 2). If Member States can demonstrate that the aforementioned preventive coexistence measures are not sufficient to achieve the desirable levels of purity, the European Commission gives those Member States the possibility “to exclude GMO cultivation from large areas of their territory” [23].

The European Coexistence Bureau (ECoB) recently issued a report on best practices to ensure the coexistence between maize cropping systems. Based on a comprehensive assessment of data generated in field experiments, commercial cultivation of GM maize, as well as modeling exercises under European environmental conditions, the report describes best management practices that can be put in place for each potential source of admixture [77]. It is important to

realize that the level of containment needed to ensure coexistence will mainly be driven by the established tolerance thresholds: the lower the tolerance threshold, the stricter the on-farm measures needed to meet labeling requirements. Apart from defining the level of containment needed, tolerance thresholds also determine the level of GM material that initiates the need to redress economic harm due to adventitious mixing. The adventitious presence of GM material above the tolerance threshold set out in EU legislation triggers the need for the product that was intended to be a non-GM product, to be labeled as containing GMOs. In other words, the product has to be labeled as containing GMOs only in cases when the established threshold is exceeded. According to the European Commission,

- “The presence of GM material in food and feed has an economic effect only when it exceeds the 0.9% labelling thresholds”, but as mentioned previously, “Member States can aim at levels of unintended GMO presence that are lower than the 0.9% labelling threshold in cases where the potential loss of income of organic and some conventional producers may be due to the presence of GMO traces at levels lower than 0.9%” [23].

In its Communication to the Council and the European Parliament from 2003, the European Commission has clearly emphasized that coexistence measures should not go beyond what is necessary to ensure that the unintentional presence of GM material in non-GM products remains below established labeling thresholds, and thus should be proportionate to the objective that is pursued. This point was reiterated in its 2010 coexistence recommendation [23], in which the European Commission stated that

- “Coexistence measures should avoid any necessary burden for farmers, seed producers, cooperatives and other operators associated with any production type”. Moreover, it argued that “the choice of measures should take into account the regional and local constraints and characteristics, such as the shape and size of fields in a region, the fragmentation and geographical dispersion of fields belonging to individual farms and regional farm management practices”.

While some Member States have taken this advice into account for most conventional producers, others

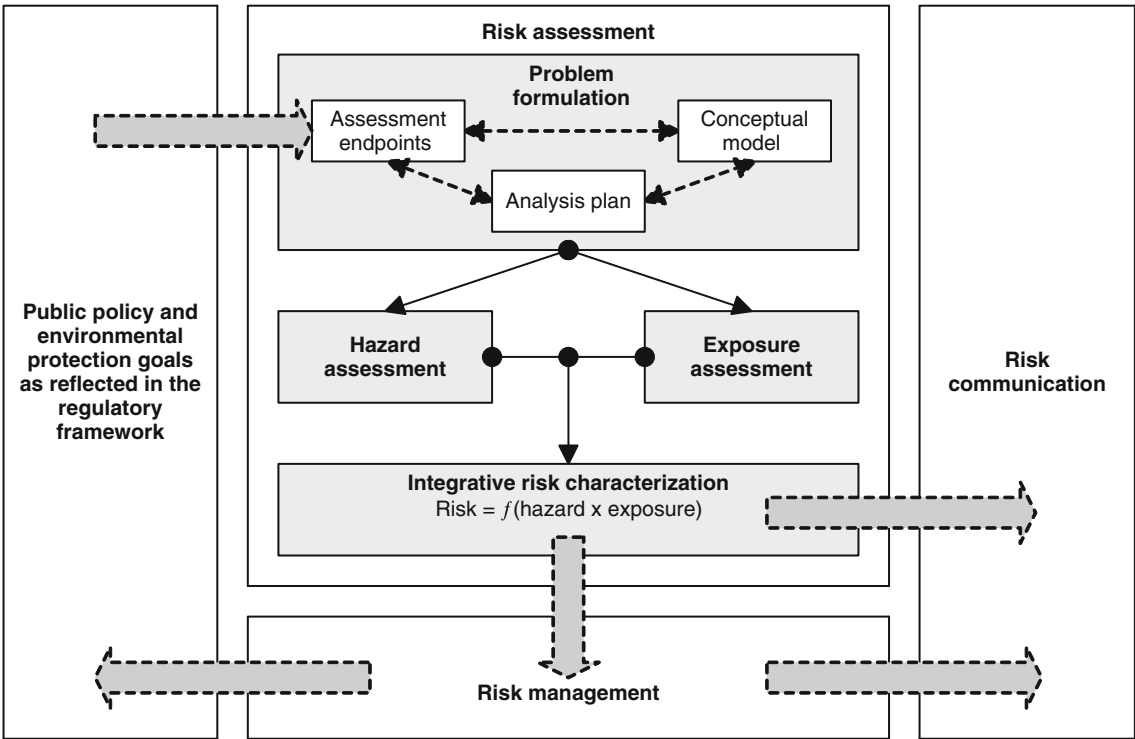
decided to propose or adopt measures that aim at keeping the amount of GM material present in conventional products as low as possible. In some cases, proposed measures, such as large and fixed isolation distances between fields of GM and conventional crops, appear to entail greater efforts for GM growers than necessary, which raises questions about the proportionality of these measures. Several authors have argued that wide and fixed isolation distances, as currently proposed by several Member States, fail to satisfy several challenges [58, 59, 78–85]. First, they are excessive from a scientific point of view; second, they are difficult to implement in practice; third, they are inconsistent with regional heterogeneity of farming; and fourth, they are not proportional to the economic incentives for coexistence. To enable regionally and economically proportionate coexistence, policy makers should allow integrating flexibility in *ex ante* coexistence regulations. This might be achieved by allowing plural coexistence measures that are adaptable to local

farming and cropping conditions, and that are negotiable amongst farmers. Computer-based decision support tools may thereby play a crucial role in the future case-by-case-based coexistence approach, as they allow the prediction of achievable levels of coexistence between neighboring maize fields under various conditions. One caveat here is that policy makers may be reluctant to adopt such a case-by-case coexistence approach because of the difficulties of implementation.

Risk Assessment Principles

Interplay of Risk Assessment, Risk Management, and Risk Communication

GMOs and their derived food and feed products are generally subject to a risk analysis before they can be commercialized [49, 86]. In the EU, the risk analysis consists of three components: risk assessment, risk management, and risk communication (Fig. 3). In risk assessment, potential adverse impacts associated with



Transgenic Crops, Risk Assessment and Regulatory Framework in the European Union. Figure 3
Risk analysis: The diagram depicts the main components of risk analysis, together with the successive steps comprising the environmental risk assessment of GM plants (Figure reprinted from [50])

a specific activity are scientifically characterized on a case-by-case basis, while in risk management, policy alternatives to accept, minimize, or reduce the characterized risks are weighed and, if needed, appropriate prevention and control options are selected. Because risk managers and regulators rely on risk assessments to make an informed decision on whether or not to approve a certain use of a GM plant, it should explain clearly what assumptions have been made during the risk assessment, and what is the nature and magnitude of uncertainties associated with the characterized risks. The decision whether a certain risk is acceptable and/or tolerable under a particular set of conditions is not part of the risk assessment itself, but part of the wider risk analysis, as this choice is not only based on scientific criteria, but also involves political, social, cultural, and economic considerations. Risk management is also functionally and temporally separate from risk assessment in order to reduce any conflict of interest and to protect the scientific integrity of risk assessment [5]. Risk communication is defined as an interactive exchange of information and opinions on risk throughout risk analysis, between risk assessors, risk managers, and other interested parties. It includes the explanation of risk assessment findings and of the basis upon which risk management decisions are made [87].

Even though there are considerable differences in regulatory requirements for GM crops between countries, environmental priorities (including the preservation of biodiversity) as well as risk terminology, most risk assessments of GM crops follow a science-based assessment process that estimates the level of risk through comparison with a non-GM counterpart [49, 88]. In addition, regulatory requirements involve the consideration of a range of issues relevant to the overall risk assessment in order to determine the impact of the GM crop on human/animal health and the environment relative to the non-GM crop, and thus its relative safety [89, 90]. Some of these elements are discussed in the next sections.

Risk Assessment Methodology and Terminology

Despite the considerable variation among risk assessment frameworks for GM plants, risk assessment generally comprises several sequential steps: (1) problem formulation as a critical first step (including hazard identification); (2) hazard characterization, that

examines potential hazards and their magnitude; (3) exposure characterization, that covers levels and likelihood of exposure; and (4) integrative risk characterization, in which the magnitude of consequences and the likelihood of occurrence are integrated (Fig. 3). In the EU, the guidance notes supplementing Annex II of Directive 2001/18/EC on the principles for the environmental risk assessment requires the definition of mitigation measures as a fifth step in the environmental risk assessment in order to manage identified risks. These mitigation measures aim to reduce the identified risks associated with GMO deployment to an acceptable level and should consider defined areas of uncertainty. Accordingly, any proposed mitigation measure is considered for the evaluation of the overall risk of GMO deployment.

Various single risk assessment studies have postulated dire risks when all they have done is characterized either a hazard associated with the use of GM crops, or an exposure to the GM crop without demonstrating whether this exposure is hazardous [5]. These studies confuse hazard or exposure with risk. In this way, they only allow speculative conclusions and contribute to uncertainties in the environmental risk assessment process. Moreover, it is important to distinguish risk from hazard, as the terms have different meanings. The term hazard is associated with the potential of an agent or situation to cause adverse effects or harm. It refers to an inherent property of that agent or situation. Risk is recognized as a function of the probability and severity (magnitude) of an adverse effect occurring to human and animal health or the environment, following exposure to a hazard under defined conditions. Without hazard, there can be no harm and thus no risk.

Problem Formulation (Including Scoping and Planning)

In order to identify the areas of greatest concern or uncertainty related to risks, each risk assessment begins with the identification and formulation of the problem, usually in the context of regulatory decision making [91]. In this respect, the most important questions to be solved and meriting detailed risk characterization are identified [92].

Problem formulation involves: (1) the identification of characteristics of the GM crop capable of

causing potential adverse effects to the environment (hazards) together with (2) the identification of pathways of exposure through which the GM crop may adversely affect the environment. This process comprises a qualitative description of the nature of potential adverse effects and its potential triggers, as well as of plausible exposure pathways to the GM crop that might result in environmental harm. Problem formulation also involves: (3) the definition of assessment endpoints, which are explicit and unambiguous targets for protection extracted from legislation and public policy goals; and (4) outlining specific hypotheses to guide the generation and evaluation of data in the subsequent successive risk assessment steps. It involves the development of a methodology – through a conceptual model and analysis plan – that will help to direct the risk characterization and to produce information that will be relevant for regulatory decision making [92–96]. In this way, problem formulation helps to make the risk assessment process comprehensive and transparent by summarizing existing scientific knowledge and explicitly stating the assumptions and principles underlying the risk assessment. Information considered in problem formulation, comprising existing knowledge of the GM crop and its interactions with the receiving environment, takes many forms. It includes published scientific literature, expert opinions, and research data. Data generated to characterize the GM crop during product development by applicants, and subsequently submitted to regulatory authorities as part of GM crop market application also provide an important source of information considered in the frame of problem formulation. Data to support problem formulation can derive from molecular, compositional, and agronomic/phenotypic analyses performed during GM crop development. Information of this type is frequently available in advance of the environmental risk assessment in order to characterize the GM crop [97–101]. If the level and quality of the available information is high, then the risk assessment can reduce the number of risk hypotheses that need to be tested for risk characterization [94, 95].

Hazard Identification: Characterization of the GM Crop Problem formulation starts with the precise characterization of the GM crop in order to identify

potential hazards. This is usually achieved through a comparative safety assessment. A comparison of the characteristics of the GM plant with those of its non-GM counterpart enables the identification of similarities and differences between the two. It enables the characterization of intended differences that were the target of the genetic modification, as well as the identification of unintended differences that may lead to harm [97–101]. The identification of differences in problem formulation is considered of primary importance, because it will direct the subsequent course of actions in the risk assessment [92, 102]. While some differences may be deemed irrelevant to the assessment, others may be meaningful in terms of posing harm to the environment or to human and animal health. Should meaningful differences be identified, the risk that they pose must be evaluated, together with those of the intended modifications, such that an evaluation of their possible biological/ecological relevance can be made (requiring them to be directly linked to those aspects of the environment that are legally protected from harm – see below).

From Protection Goals to Assessment Endpoints A crucial step in problem formulation is to identify aspects of the environment (e.g., valued entity, ecosystem service/function) that need to be protected from harm according to protection goals set out in existing EU legislation, and to translate those into concrete measurable phenomena. Risk assessors need to know *what* to protect, *where* to protect it, and over *what time period*. Usually, aspects of the environment to be protected can be divided into two discrete but interconnected categories: the protection of biodiversity (biodiversity conservation), and the protection of functions/services provided by ecosystems. Ecosystem *services* are distinct from ecosystem *functions* by virtue of the fact that humans, rather than other species, benefit directly from these natural assets and processes. The benefits to humans are many and varied; however, those ecosystem services relevant in an agricultural context include food and feed production, pollination, pest regulation, decomposition of organic matter, soil nutrient cycling, soil structure and formation, water regulation and purification, and cultural services (such as aesthetic value and recreation). The precise relationship between biotic and abiotic processes and the ecosystem functions they drive

is an area of considerable scientific debate. Some species contribute to ecosystem functioning in ways that are unique and hence their addition or loss from a community would cause detectable changes in functioning. Most ecosystems, however, exhibit functional redundancy, where several species are able to perform the same critical function. These species are at least partly substitutable and their loss can be compensated for by other species.

The challenge to problem formulation is that, in EU legislation, protection goals are general concepts that are defined in broad terms, which often are too vague to be scientifically useful in the frame of environmental risk assessments of GM crops. Moreover, a broad array and diversity of protection goals are mentioned in EU legislation. The resultant interpretation is that everything should be protected, everywhere, all of the time. However, to be useful in terms of ecological risks and decision making, it is important that general and broadly defined protection goals are translated into concise and concrete measurable assessment endpoints. If the protection of biodiversity is a public policy goal, then a typical assessment endpoint can be the abundance and species richness of certain groups of organisms at a relevant life stage within a landscape or region over a specific time period [102]. Assessment endpoints represent aspects of the environment to be protected and that can be assessed: they are defined by the valued attribute and/or function (e.g., control of arthropod pest populations) of an ecological entity (e.g., arthropod natural enemies) that could be affected adversely by the GM crop and that requires protection from harm [103]. It is not an abstract goal such as ecosystem health or sustainability, but a real, operationally definable property of an aspect of the environment that reflects management or protection goals laid down in EU legislation. Because arthropod natural enemies fulfill relevant ecological services by contributing to the natural regulation of arthropod pest populations within crop fields, they can be identified as the entity to be preserved; similarly, the biological control functions they perform can be identified as the attribute/function (e.g., [104]). It is important that assessment endpoints are defined as far as possible using measurable criteria relevant to the case under study, such that any change in these endpoints can be readily identified. Moreover, assessment endpoints

should ideally be selected based on: relevancy to the protection goal; ecological relevance; susceptibility to the potential stressor; and practicality [105].

Once assessment endpoints have been set, the level of environmental protection to be maintained, or the environmental quality to be preserved, needs to be defined. This process includes establishing the *harmful effect* and both the *spatial* and the *temporal scales* relevant for the ecological entity and its attribute/function to be preserved (Table 1). The harmful effect describes to what extent the environmental quality should be preserved (or above what threshold a difference between the GM crop and its appropriate comparator may lead to harm and would be considered a disturbance in environmental quality). For the conservation of biodiversity, a relevant decrease in abundance of protected or valued species can be seen as a harmful change. Similarly, for ecosystem services, a relevant disturbance in ecological function can be seen as a harmful change [106]. The spatial and temporal scales are the habitats in which, and the period during which, environmental quality should be preserved, respectively [104, 107].

Conceptual Model: Exposure Profiles and Hypotheses

The conceptual model describes the consequential exposure scenario of how harm to the assessment endpoint may arise from GM crop deployment and allows for a characterization of risks. Key relationships are described: between the GM crop and the valued entity to be protected from harm; pathways of exposure through which the GM crop may affect the valued entity either directly or indirectly (= exposure profile); and any potential impact of the GM crop to the environment [92, 96]. The conceptual model includes available information on: the nature of the stressor; its intended uses; reasonable exposure profiles; and potential responses of the assessment endpoint as a result of exposure. It can take an array of forms, from the simplest of statements to complex flowcharts and diagrams.

Environmental pathways and levels of exposure will vary depending upon the intended uses of a GM crop, such as import, processing, food, feed, and/or cultivation. In the case where the use of GM plants does not include cultivation in the EU, problem formulation will consider exposure: (1) via accidental release of propagules, such as seeds of the imported commodity spilled into the

Transgenic Crops, Risk Assessment and Regulatory Framework in the European Union. Table 1 Criteria to operationalise protection goals^a

Criterion	Measurable variable
Ecological entity	Species of conservation concern; ecosystem service
Attribute	Abundance; ecological function
Unit of protection	Individual; population; community; functional group; assemblage; guild
Harmful effect	Relevant decrease in abundance; relevant disturbance in ecological function
Spatial scale of effect	Field; field margin; other agricultural land; non-agricultural habitats; landscape
Temporal scale of effect	Generations; days; weeks; months; seasons; years

^aTable adapted from the stakeholders' workshop "Protection goals for environmental risk assessment of pesticides: what and where to protect?" organized by the EFSA Panel on Plant Protection Products and their Residues (PPR) (<http://www.efsa.europa.eu/en/scdocs/scdoc/1672.htm>); see also [108, 134]

environment during transportation and processing, potentially leading to sporadic feral GM plants; (2) indirect exposure through manure and feces from the gastrointestinal tracts of animals mainly fed the GM crop; and/or (3) manure or organic plant matter either imported as a fertilizer or soil amendment or derived from other byproducts of industrial processes. In the case where the GM crop use includes cultivation in the EU, problem formulation will consider exposure resulting from the expected cultivation in the receiving environment.

A well-structured conceptual model in which the components of the system are detailed will allow the identification and formulation of relevant hypotheses that arise from the consideration of potentially significant risks. These hypotheses are necessary to make assumptions and predictions about how a stressor or identified difference in the GM plant could affect and pose harm to an assessment endpoint [94, 95]. Importantly, within the analysis phase, hypotheses amenable to testing and corroboration are defined [92] which further guide the methodological approach taken to evaluate the magnitude of harm [92, 102].

Analysis Plan: Statistical Considerations and Experimental Design

The last step of problem formulation comprises an analysis plan in which decisions are made about the most appropriate way to measure the response of each assessment endpoint to GM crop deployment. In this planning phase, approaches are delineated for the generation and evaluation of requisite data, in order to test hypotheses formulated in the conceptual model. Reasonable scenarios are placed in an analysis plan by describing and selecting: (1) the various measures to be used (measurement endpoints) in the assessment and subsequent risk characterization; and (2) the description of methods and criteria for measurement.

The measurement endpoints, derived directly from the assessment endpoints, define the indicator of change that will actually be recorded as part of a comparative risk assessment study [107], and usually constitute estimates of *exposure* or *hazard*. Measures of exposure cover properties of the GM crop and are described in terms of the route, frequency, duration, and intensity of exposure relative to the valued entity [92], while measures of hazard represent the measurable change to the valued entity in response to a changed attribute (e.g., transgenic protein) of the GM crop to which it is exposed [107]. Measures of hazard may be an acute lethal concentration resulting in the death of 50% of the organisms tested (LC₅₀), or a chronic *no observable adverse effect level* (NOAEL) measured for the valued entity [92]. How exposure measurement relates to the hazard measurement is described in the risk formulation.

Once specific measurement endpoints are chosen and given a priority, appropriate methods and criteria of measurement are selected and described in the analysis plan [93]. This includes information on: studies to be conducted; the appropriate tier for analysis; the design of protocols; and statistical power [96, 102, 106, 107, 109, 110]. The selection and prioritizing of measures to be used and testing needed enable the allocation of appropriate human and financial resources [111], so that only essential data for risk characterization are collected [94].

It is important to realize that for practical reasons not all potentially exposed non-target organisms can be tested for regulatory purposes [102]. Therefore, it is necessary to select an appropriate subset of species that

can be tested effectively under laboratory conditions [102], or that are available in sufficient numbers in the field to give statistically meaningful results [96, 112–115]. This selection of species is based on several criteria, comprising: ecological relevance of the species; the likely exposure of the species to the GM crop; species susceptibility to known or potential stressors; the anthropocentric value of the species; and species testability [96, 102, 113, 116]. The environmental risk assessment may also consider species with special aesthetic or cultural values, or species of conservation importance and that are classified as threatened or endangered, although these are unlikely to be tested directly but substituted by a surrogate species in tests. The number and type of species that are to be tested will depend upon the hypotheses generated during the conceptual model.

Once assessment and measurement endpoints have been set and approaches have been delineated for the generation and evaluation of requisite data to test hypotheses formulated during problem formulation, the level of environmental change that constitutes harm is to be set through *limits of concern*. These limits of concern define the minimum relevant ecological effect that is deemed of sufficient magnitude to cause harm, and is therefore biologically significant. Limits of concern are directly related to the type of studies that are to be performed either in the laboratory or in the (semi)-field. For laboratory studies, limits of concern are conservative trigger values which, if exceeded, will indicate potential risks and the need for exposure assessments and determination of field scale effects. For field studies, the lower limit, which corresponds for example to a decrease in the abundance of a particular species in the presence of the GM plant relative to that for its appropriate comparator, will usually be defined by the threshold effect deemed to be of just sufficient magnitude to cause environmental harm [106]. Knowing in advance what effect size is to be determined is crucial, as this information will enable to design the study in a way that it has sufficient statistical power to detect the anticipated effect. Limits of concern are estimated from literature data, modeling, and existing knowledge, and may be based on political, social, cultural and economic considerations. To define the minimum relevant ecological effect that is deemed biologically significant and that is deemed of

sufficient magnitude to cause harm, it is important that (sets of) limits of concern for each assessment endpoint are set. Usually, the lower limit, which corresponds for example to a decrease in the abundance of a particular species in the presence of the GM crop relative to that for the non-GM counterpart, will be defined by the threshold effect that was deemed to be of just sufficient magnitude to cause environmental harm [106]. If these limits are exceeded, then detailed quantitative modeling of exposure may be required to scale up effects at the field level, both temporally and spatially. For studies in environment(s) that are controlled, the limits of concern will usually be trigger values which, if exceeded, will either lead to conclusions on risks or the need for further assessment in field [106]. Limits of concern can be defined by, for example, literature data, modeling, existing knowledge and policy goals. Ideally, these limits should be set by risk managers, as they will describe the extent of impact tolerated on a protection goal resulting from GM crop market approvals.

Conclusion Problem formulation is the crucial starting point for risk assessments, as it enables a structured, logical approach to detecting potential risks. Having a properly constructed analysis plan based on a conceptual model that is clearly linked to assessment endpoints helps to guide the collection of data that are relevant to investigating the possible safety of a GM crop. Moreover, it helps the risk assessment process to be comprehensive (by summarizing existing knowledge of the system under study) and transparent (by explicitly stating significant assumptions underlying the risk assessment, and ultimately regulatory decision making). In contrast, poor problem formulation can lead to the collection of data that are either unnecessary, superfluous, or irrelevant, diverting time and efforts from the more serious of the identified risks, thereby slowing down the procedure and increasing associated costs [90, 92, 94, 95, 117].

Risk Assessment Concepts and Approaches

Several concepts and approaches are considered during the risk assessment of GM crops. Risk assessment of GM crops: (1) is science-based, where quantitative information is available, and uses qualitative

information in the form of expert judgment; (2) uses a comparative approach; (3) is case specific; (4) is iterative and, in a transparent manner, examines previous conclusions in light of new information; and (5) follows a tiered approach.

Science-Based Assessment

The evaluation of potential adverse effects is based on scientific and technical data, and on common methodology for the identification, gathering, and interpretation of relevant data.

Comparative Safety Assessment and Familiarity Concept

The risk assessment strategy for GM crops seeks to use appropriate methods to compare a GM crop with its non-GM counterpart. The importance of risks posed by a GM crop is placed in the context of risks posed by its non-GM counterpart. A two-step approach is followed, starting with the identification of possible differences between the GM and non-GM counterpart (= proof of difference); which is then followed by the assessment of the environmental and food/feed consequences of any identified differences (= proof of equivalence) (see [106] for further details). The proof of difference approach verifies whether the GM crop is different from its non-GM counterpart, and may lead to the characterization of a potential risk depending on the type of the identified difference, and the extent and pattern of its exposure. The proof of equivalence approach aims to verify whether the GM crop is equivalent to its non-GM counterpart within bounds defined by so-called limits of concern. The two approaches are complementary: statistically significant differences may point to biological changes caused by the genetic modification, but these may or may not be relevant from the viewpoint of environmental harm [106, 118].

For the *food/feed safety assessment*, the equivalence test involves a comparison of the GM crop with a selection of non-GM counterparts, such as commercial varieties, aiming to identify any values of the GM crop that lie within the range of values from natural variation (i.e., observed among the set of non-GM counterparts included in the trial). The underlying assumption is that traditionally bred

plants have a history of safe use for the consumer or animals and the environment, and familiarity for the consumer. For the *environmental risk assessment*, conventional cropping systems are not characterized as having a history of safe use, therefore, a near-isogenic conventional counterpart is used to identify any differences with the GM crop which might result in environmental harm. However, in any agricultural system, environmental impact critically depends upon crop management practices, such as the application of pesticides, environmental stewardship, and any mitigation measures adopted. Hence, the characterization of management regimes is an integral part of the definition of whatever comparator is chosen for the environmental risk assessment of GM crops [106].

The concept of familiarity is based on the fact that most GM crops are developed from more traditionally bred crops, the biology of which is well known. The knowledge about the non-GM crop, gained through experience over time, can therefore be used in a risk assessment to establish differences associated with the genetic modification and the subsequent management of the GM crop. According to the Organisation for Economic Co-operation and Development (OECD), familiarity is derived from the knowledge and experience available from conducting a risk analysis prior to the scale-up of any new crop variety in a particular environment, and from previous GM crop market applications for similar constructs and traits in similar or different crops [119]. However, it is important to bear in mind that familiarity is not an endpoint in risk assessment and does not necessarily equate to safety. If differences between the GM crop and its appropriate comparator have been identified, it needs to be determined whether these differences have any significance for the assessment endpoint under consideration.

The comparative safety assessment is usually based on data from three different sources: molecular, compositional, and agronomic/phenotypic analyses. The molecular characterization includes a precise characterization of the inserted DNA, information on which genes are expressed in the GM crop, and evidence that no detectable unintended effects have occurred because of the insertion [86]. Agronomic, phenotypic, and compositional data collected from field trials carried

out in a range of agricultural environments that are typical of the place where the crop is grown will highlight whether the GM crop is equivalent in its functional properties to the non-GM counterpart and, in particular, how stable they are from year to year and across environments.

Molecular Characterization In an environmental risk assessment, molecular characterization data can be used to test the hypothesis that the inserted DNA does not disrupt endogenous plant genes, or trigger the production of new proteins/metabolites (other than the intended ones). If any endogenous plant genes were disrupted, rearranged, or overexpressed, these differences would be considered as unintended effects of the genetic modification, and would be further assessed to determine the biological significance thereof. A variety of data and information is therefore necessary as no individual test can detect all possible unintended effects or identify, with certainty, those relevant to human health or environmental impact. Specifically, detailed information is usually required on: the source of the donor DNA; the transformation method; the organization of the inserted DNA at the insertion site(s); and on the expression and stability of the insert. Information concerning vector construction and the transformation method employed, as well as on all known gene regulatory elements and coding sequences is also relevant. Many authorities require the modification system to be mapped to a degree consistent with available technology, usually to the level of base sequence. An evaluation of whether the transgenes behave similarly to endogenous plant genes in their stability and inheritance between generations is also carried out. Hence, a risk assessment would have to be conducted to determine whether any detected unintended effects could result in potential adverse effects on human or animal health and the environment [100, 101]. Notably, insertional effects (intended or otherwise) are related to the specific transformation event, and not to the function of the inserted transgene; therefore, as with random mutation, most unintended changes caused by deliberate insertions of genetic material will be, at best, neutral to the plant. It is extremely rare for a significant improvement in ecological fitness or substantial compositional changes to result [120].

Compositional Analysis Compositional analyses enable the determination of any differences in key components as a result of the genetic modification, and to test the hypothesis that no difference in key component concentrations exists between the GM crop and its non-GM counterpart. Compositional data are generated in field trials carried out in several locations throughout the intended area of cultivation of the crop. In these field trials, the GM crop is grown alongside its appropriate comparator, and samples are taken. Parameters are selected that are typical for the crop that is assessed, and that are representative of the main metabolic pathways. Significant changes in these parameters are expected to be indicative of more fundamental changes in the crop that requires further evaluation [121]. The lack of statistically significant differences in the concentration of multiple analytes is a strong indication that the genetic modification has not introduced new constituents, nor harmful unintended changes. However, should statistically significant differences be detected, a harmful change is not necessarily indicated: should the different concentration of a specific analyte in the GM crop be within the range known for the crop, then it can be considered biologically insignificant, otherwise they will constitute “unintended identified” differences that require consideration in the risk assessment [100, 101].

Agronomic and Phenotypic Characterization Information on agronomic and phenotypic characteristics is obtained from multi-location agronomic field trials representative of different environments where the GM crop may be grown. To assess the agronomic performance of the GM crop, a sound understanding of the biology of the crop is required, as well as its relationship with the environment in which it is to be released. A variety of plant characteristics, such as plant vigor, growth habit, yield, crop quality, insect and disease susceptibility, fertility, dispersal mechanisms, and endogenous toxins, are recorded and any differences in the GM crop identified, which could potentially cause it to become a weed of agricultural or natural habitats, or otherwise interact differently than the non-GM counterpart in the environment. If there are no significant differences in characteristics associated with survival, growth, and reproduction, then it is likely that the genetic modification did not

alter the persistence, invasiveness or gene flow potential of the GM crop. If significant differences are identified, the values obtained in the GM crop will be assessed for their biological relevance by comparing them with the range of values known for (non-GM) commercial varieties [98–101].

Case-by-Case Principle

No two environmental risk assessments are the same, as each is dependent upon a range of variables found when considering: both the source and target environments; the biological and ecological characteristics of the GM crop; the scale and frequency of the proposed deliberate release; and the interactions amongst them [97, 122]. It is obvious that each range of variables will also differ between risk assessments, and thus a case-by-case approach is taken. In this way, the required information for the environmental risk assessment may vary depending on: the plant species under consideration; its intended use; and potential receiving environments, taking into account specific cultivation requirements and the presence of GM crops already in the environment.

Iterative and Adaptive

It is recognized that an environmental risk assessment is framed within the scientific knowledge available at the time it is conducted, and that regulatory decisions are made in this context. The environmental risk assessment has to take into account uncertainty at various levels, and which may arise from: limitations in the data (e.g., limited exposure data); gaps in the effect database; the limitation of the test systems and measurement endpoints selected; inadequacy of study designs; and the uncertainties in extrapolating between species. Scientific uncertainty may also arise from differing interpretations of existing data or the lack of some relevant data. Uncertainty may relate to qualitative or quantitative elements of the analysis [106]. Although it may be impossible to identify all the uncertainties present, the assessment should include a description of the types of uncertainties encountered and considered during the different risk assessment steps. Their relative importance and influence on the assessment outcome should also be

described. Under current EU legislation, a high precision in environmental risk assessments becomes near impossible to achieve, as the identification of any areas of uncertainty which relate to areas outside current knowledge and the limited scope of the assessment is a mandatory requirement, for example, the impact of large-scale exposure of different environments from GM crop cultivation, the impact of exposure over long periods of time, and any cumulative long-term effects.

PMEM, which became mandatory under current EU legislation, aims to identify possible unanticipated adverse effects on human health or the environment which could arise directly or indirectly from GM crops. It also allows for the collection of additional data during the cultivation phase. The scientific knowledge obtained during the monitoring of GM crops, along with experiences gained from their cultivation and any other new knowledge (generated through biosafety research), provides valuable information to risk assessors to update environmental risk assessments and reduce any remaining uncertainties. In the EU, the objectives of a monitoring plan according to Annex VII of Directive 2001/18/EC are: (1) to confirm that any assumption regarding the occurrence and impact of potential adverse effects of the GMO, or its use, in the environmental risk assessment are correct; and (2) to identify the occurrence of adverse effects of the GMO, or its use, on human or animal health or the environment which were not anticipated in the environmental risk assessment. A PMEM plan of GM crops is mandatory in all market approval dossiers submitted under Directive 2001/18/EC and the GM Food and Feed Regulation.

In the EU, PMEM is composed of *case-specific monitoring* and *general surveillance*. Case-specific monitoring is not obligatory, but may be required to verify risk assessment assumptions and conclusions, whereas a general surveillance plan, as part of a GM crop market application, is a legal obligation. Due to different objectives between case-specific monitoring and general surveillance, their underlying concepts differ [123]. Case-specific monitoring enables the determination of whether, and to what extent, anticipated adverse effects occur during GM crop deployment, and thus to relate observed changes to specific causes. It is mainly triggered by scientific uncertainties that were identified

in the environmental risk assessment. Therefore, a hypothesis is established that can be tested on the basis of newly collected monitoring data (“bottom-up approach”). In general surveillance, by contrast, the general status of the environment that is associated with the GM crop deployment is monitored without any preconceived hypothesis in order to detect any possible effects that were not anticipated in the environmental risk assessment, or that are long term and cumulative. Should any such effects be observed, they are studied in more detail to determine whether the effect is adverse and whether it is associated with the use of a GM crop [87]. General surveillance data can originate from various sources: (1) farm questionnaires; (2) existing surveillance networks (such as plant health surveys, soil surveys, ecological and environmental observations); (3) scientific literature; (4) industry stewardship programs; and (5) alert issues. Questionnaires for farmers form a useful part of general surveillance, as this tool enables the reporting of any observations of effects linked with GM crop cultivation: farm questionnaires use first-hand observations and rely on farmers’ knowledge and experience of their local agricultural environments, comparative crop performance and other factors that may influence events on their land [124]. With this tool, monitoring focuses mainly on the cultivation area of GM crops and its surroundings, which is relevant to protection goals such as sustainable agriculture, soil function, or plant health. Aspects of biodiversity, however, are only addressed indirectly (such as the adoption of conservation/no-till practices, rotation regimes, biological control failures) and may not be sufficiently resolved [125]. Therefore, additional sources of information should be considered in the frame of general surveillance. Moreover, farm questionnaires should include different combinations and mixtures of GM crops that may be grown in sequence or rotation in fields [126]. Directive 2001/18/EC proposes to make use of established routine surveillance networks. EU Member States have various networks in place – some of which have a long history of data collection – that may be helpful in the context of general surveillance of GM crops. The networks involved in routine surveillance offer recognized expertise in a specific domain and have the tools to capture information on important environmental aspects over a large geographical area.

However, networks fully meeting all the needs of GM crop monitoring are limited: most do not always provide data of relevance to monitoring the impacts of GM crop cultivation [125, 127]. Concern has been raised, predominantly outside of the EU, however, that in the absence of a valid hypothesis, PMEM for undefined hypothetical adverse effects from a GM crop is not feasible, and adds nothing to the premarket testing results, while potentially undermining confidence in the overall safety assessment process [128].

Risk assessments undertaken in the EU are always iterative, in the sense that regulatory decisions are temporary, reversible, and adaptable in light of new information that becomes available. Under Directive 2001/18/EC, the requirement of reexamination has been strengthened by limiting the duration of any market consent to a maximum period of ten years.

Tiered Approach

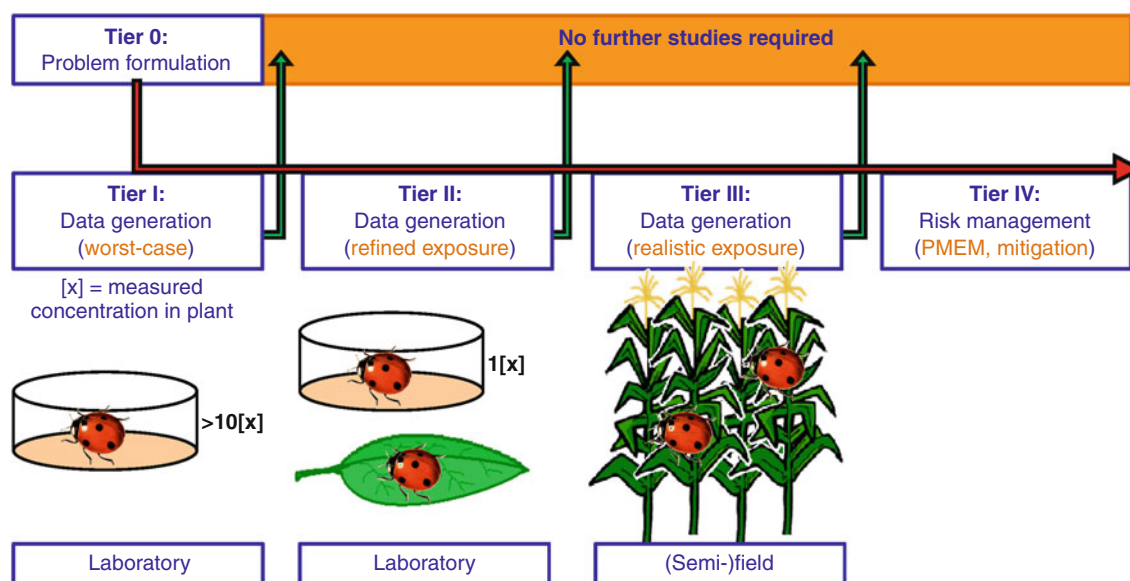
An environmental risk assessment is generally conducted in a tiered manner, where information collected in lower tiers directs the extent and nature of any experimentation conducted in higher tiers (Fig. 4). Thereby, hazards are evaluated within different tiers that progress from worst-case scenario conditions, framed in highly controlled laboratory environments, to more realistic conditions in the field [87, 97, 102, 122, 129–131]. Usually, three tiers is the norm, comprising of experimental tests under controlled conditions (tier 1: laboratory tests where test organisms are exposed to a high level of the newly expressed protein as pure protein; tier 2: laboratory tests where test organisms are exposed to refined, more realistic, levels of the newly expressed protein as pure protein or plant material); and (semi-)field tests (tier 3). Within each tier, all relevant data are gathered in order to determine whether there is sufficient information to conclude the risk assessment at that tier. The conclusion can only be made if any residual uncertainty has been defined, otherwise additional investigations to generate further data at a higher tier (s) are designed and undertaken to do so [102]. In the case that no reliable risk conclusions can be drawn, decision making can consider whether risk management measures (tier 4) should be put in place to reduce the overall risk. It is important that throughout the

assessment, the problem being addressed (tier 0) remains appropriate and is revised, if necessary.

Lower-tier tests serve to identify and test potential hazards under worst-case scenario conditions and thus involve conservative assumptions, acknowledging that the likelihood of detecting any potential adverse effects of the GM crop or its products on target and nontarget biota increases directly with increasing (usually excessive) levels of exposure. These studies are conducted under controlled laboratory or growth room conditions in order to: quantify effects in relation to known exposure levels; to provide high levels of replication and control; and to increase the statistical power for testing the established hypotheses. Effects of the GM crop on indirectly exposed organisms, that is, those that are one or two steps removed in the food chain (e.g., predators and parasites of primary phytophagous or plant pathogenic organisms), are generally assessed in the second tier, which is also generally conducted under controlled laboratory, growth room or glass-house conditions in order to measure effects in relation to known exposure levels [87]. If no hazards are identified and the GM crop is not different from the appropriate comparator, the tested product may be regarded

as safe, and no further testing at a higher tier deemed necessary.

However, should potential hazards be detected in early-tier tests or if unacceptable uncertainties concerning possible hazards remain, additional information is required to confirm whether the observed effect might still be detected at more realistic rates and routes of exposure [87, 97, 102, 131]. Progression to larger-scale experiments in higher tiers aims to provide increasingly refined estimates of exposure. Field trials are then established in which the cultivation of the GM crop is conducted with greater environmental realism and thus relevance. As such, actual levels of exposure of different biota can be quantified. In comparison with the appropriate comparator and its management, likely ecological adverse effects due to the GM plant and its management can be determined. While higher-tier studies offer greater environmental relevance, they may have lower statistical power due to the higher variability of environmental conditions (e.g., climate) that can mask effects generated by the GM crop or its product [131]. In exceptional cases, higher-tier studies may be conducted at an initial stage when early-tier tests are not possible or



Transgenic Crops, Risk Assessment and Regulatory Framework in the European Union. Figure 4
Tiered approach to non-target organism testing (Figure adapted from [97])

meaningful. As such, many risk assessments are conducted in a tiered manner, meaning that risk assessment studies increase in complexity depending upon the findings at each level of assessment [91]. In cases where uncertainty about the risk remains after higher-tier studies, one can always return to lower tiers to conduct additional studies [102].

The tiered approach is consistent with the iterative or adaptive nature of risk assessment where conclusions are reviewed when new information is obtained. Uncertainty in risk assessment is reduced because each tier is guided by results obtained in the previous tier, and specific, testable, and relevant hypotheses are formulated based on these data [87, 97, 102, 122, 131].

Future Directions

Since one of the main objectives of the EU GMO regulatory framework is to ensure a high level of protection of human and animal health and the environment, the focus is on the assessment of risks. Whether GM crops fulfill wider socioeconomic and ecological aspirations is not considered explicitly. However, when analyzing potential risks, it is important to bear in mind that the real choice is not between GM crops that are inherently risky and traditionally bred ones that are completely safe. Both existing crops and those with novel traits (including GM crops) will have both positive and negative attributes [56]. To fully acknowledge the overall consequences of adopting specific crops, and to assess and manage more effectively the environmental footprint of agriculture as a whole, it has been suggested that broader and more balanced legislative oversight is needed in the EU [132, 133]. At the EFSA scientific colloquium on challenges and approaches for the environmental risk assessment of GM plants [133], the discussion group on broadening the scope of the environmental risk assessment provided the following recommendations:

- A paradigm shift would be required to change from risk assessment as it is currently practiced, to a more sophisticated assessment which balances risks and benefits: (i) The focus on only GM crops defies scientific evidence. In the longer term, risk assessors could develop an alternative approach on a scientific basis. 'Novelty' is one option. (ii) The status quo, in which risk

assessment is interpreted very narrowly in terms of adverse impacts, is not sustainable, and perceptions of the quality of environmental risk assessments suffer as a result. A framework for the future is required. (iii) There is a need to build decision support tools for the risk assessors to better consider impacts of whole farming systems.

The assessment of GM crops has intersected with a wider debate about sustainable agriculture in the EU, blurring any distinctions between environmental, agricultural, and socioeconomic issues. A sustainability assessment may be helpful in recovering public and market confidence, as it integrates larger societal concerns. It enables to define and integrate the underlying values at stake, trade possible risks against benefits, compare technological alternatives, evaluate the usefulness of GM crops, and to assess a whole agricultural system. In this way, it may promote finding a better balance between agricultural production and biodiversity, and evolve toward a socially more robust risk analysis, in which sustainable development is the goal.

While such an integral sustainability assessment certainly will not make things easier and might be an unachievable ideal, it may offer a helpful framing of the debate about multiple controversial aspects. Social, ecological, and economic considerations (such as ensuring social equity and cohesion, a high level of environmental protection, and economic prosperity) that have too often been treated in isolation could then be brought together and explored in a holistic and long-term perspective. Preferences for certain options would no longer be advanced independently of the broader, more complex agricultural and societal context. Likewise, underlying values at stake and normative perspectives fueling the debate might be brought to light.

Disclaimer

Opinions and views expressed in this entry are strictly those of the authors, and may not necessarily represent those of the organizations where the authors are currently employed.

Acknowledgments

The authors wish to thank Salvatore Arpaia, Detlef Bartsch, Adinda De Schrijver, Matty Demont, Achim Gathmann, Rosemary Hails, Jozsef Kiss, Antoine

Messéan, Karin Nienstedt, Joe Perry, Dirk Reheul, Olivier Sanvido, and Jeremy Sweet for inspiring discussions that helped to develop this entry.

Bibliography

- EC (2008) The Economics of Ecosystems and Biodiversity – Interim Report 2008. <http://ec.europa.eu/environment/nature/biodiversity/economics>
- EC (2001) Directive 2001/18/EC of the European Parliament and of the Council of 12 March 2001 on the deliberate release into the environment of genetically modified organisms and repealing Council Directive 90/220/EEC. Off J Eur Communities L106:1–39
- EC (2004) Directive 2004/35/EC of the European Parliament and of the Council of 21 April 2004 on environmental liability with regard to the prevention and remedying of environmental damage. Off J Eur Communities L143:56–75
- Secretariat of the CBD (2000) Cartagena protocol on biosafety to the convention on biological diversity: text and annexes. Secretariat of the Convention on Biological Diversity (CBD), Montreal, Canada. <http://www.cbd.int/doc/legal/cartagena-protocol-en.pdf>
- Johnson KL, Raybould AF, Hudson MD, Poppy GM (2007) How does scientific risk assessment of GM crops fit within the wider risk analysis? Trends Plant Sci 12:1–5
- James C (2010) Global status of commercialized biotech/GM crops: 2010. Highlights of ISAAA briefs No 42, Ithaca, New York
- Chapotin SM, Wolt JD (2007) Genetically modified crops for the bioeconomy: meeting public and regulatory expectations. Transgenic Res 16:675–688
- Herring RJ (2008) Opposition to transgenic technologies: ideology, interests and collective action frames. Nat Rev 9:458–463
- Levidow L, Carr S, Wield D (2005) European Union regulation of agribiotechnology: precautionary links between science, expertise and policy. Sci Public Policy 32:261–276
- Devos Y, Reheul D, De Waele D, Van Speybroeck L (2006) The interplay between societal concerns and the regulatory frame on GM crops in the European Union. Environ Biosaf Res 5:127–149
- Devos Y, Maesele P, Reheul D, Van Speybroeck L, De Waele D (2008) Ethics in the societal debate on genetically modified organisms: a (re)quest for *Sense* and *Sensibility*. J Agric Environ Ethics 21:29–61
- Levidow L, Carr S (2007) GM crops on trial: technological development as a real world experiment. Futures 39:408–431
- Lofstedt RE (2006) How can we make food risk communication better: where are we and where are we going? J Risk Res 9:869–890
- Marks LA, Kalaitzandonakes N, Wilkins L, Zakharova L (2007) Mass media framing of biotechnology news. Public Underst Sci 16:183–203
- Maesele PA, Schuurman D (2008) Biotechnology and the popular press in Northern Belgium. Sci Commun 29:435–471
- Kasperson RE, Kasperson JX (1996) The social amplification and attenuation of risk. Ann Am Acad Polit Soc Sci 545:95–105
- Winickoff D, Jasanoff S, Busch L, Grove-White R, Wynne B (2005) Adjudicating the GM food wars: science, risk, and democracy in world trade law. Yale J Int Law 30:81–123
- Kuiper HA, Davies HV (2010) The SAFE FOODS risk analysis framework suitable for GMOs? Food Control. doi:10.1016/j.foodcont.2010.02.011
- Gaskell G, Allansdottir A, Allum N, Corchero C, Fischler C, Hampel J, Jackson J, Kronberger N, Mejlgaard N, Revuelta G, Schreiner C, Stares S, Torgersen H, Wagner W (2006) Europeans and Biotechnology in 2005: Patterns and Trends, Eurobarometer 64.3. http://ec.europa.eu/research/press/2006/pdf/pr1906_eb_64_3_final_report-may2006_en.pdf
- Christiansen T, Polak J (2009) Comitology between political decision-making and technocratic governance: regulating GMOs in the European Union. Eipascopie Bull 1:5–11
- EC (2010) Proposal for a Regulation of the European Parliament and of the Council amending Directive 2001/18/EC as regards the possibility for the Member States to restrict or prohibit the cultivation of GMOs in their territory. http://ec.europa.eu/food/food/biotechnology/docs/proposal_en.pdf
- EC (2010) Communication from the Commission to the European Parliament, the Council, the Economics and Social Committee and the Committee of the Regions on the freedom for Member States to decide on the cultivation of genetically modified crops. http://ec.europa.eu/food/food/biotechnology/docs/communication_en.pdf
- EC (2010) Commission Recommendation of 13 July 2010 on guidelines for the development of national co-existence measures to avoid the unintended presence of GMOs in conventional and organic crops. http://ec.europa.eu/food/food/biotechnology/docs/new_recommendation_en.pdf
- Chipman A (2010) Fears over Europe's GM crop plan. Nature 466:542–543
- Ryffel GU (2010) Making the most of GM potatoes. Nat Biotechnol 28:318
- Secretariat of the CBD (1992) Convention on biological diversity. Secretariat of the Convention on Biological Diversity (CBD), Montreal, Canada. <http://www.biodiv.org/doc/legal/cbd-en.pdf>
- James C (2009) Global status of commercialized biotech/GM crops: 2009. ISAAA Brief No. 41. ISAAA, Ithaca
- Jaffe G (2004) Regulating transgenic crops: a comparative analysis of different regulatory processes. Transgenic Res 13:5–19
- FAO/WHO (2000) Safety aspects of genetically modified foods of plant origin. A joint FAO/WHO consultation on foods derived from biotechnology, Geneva, Switzerland, 29 May–2 June 2000. World Health Organization, Geneva
- Secretariat of the IPPC (2005) Pest risk analysis for quarantine pests, including analysis of environmental risks and living modified organisms. International Standards for Phytosanitary

- Measures (ISPM) No. 11. Secretariat of the International Plant Protection Convention (IPPC), Food and Agriculture Organization of the United Nations, Rome, Italy. <ftp://ftp.fao.org/docrep/fao/008/y5874e/y5874e00.pdf>
31. Mackenzie R, Burhenne-Guilmin F, La Vina AGM, Werksman JD (2003) An explanatory guide to the Cartagena protocol on biosafety. In cooperation with Ascencio A, Kinderlerer J, Kummer K, Tapper R. IUCN, Gland, Switzerland and Cambridge, UK. www.cbd.int/doc/books/2003/B-01669.pdf
 32. Breyer D, Herman P, Brandenburger A, Gheysen G, Remaut E, Soumillion P, Van Doorselaere J, Custers R, Pauwels K, Sneyers M, Reheul D (2009) Commentary: Genetic modification through oligonucleotide-mediated mutagenesis. A GMO regulatory challenge? *Environ Biosaf Res* 8:57–64
 33. Macdonald P, Yarrow S (2003) Regulation of Bt crops in Canada. *J Invertebr Pathol* 83:93–99
 34. McHughen A, Smyth S (2008) US regulatory system for genetically modified [genetically modified organisms (GMO), rDNA or transgenic] crop cultivars. *Plant Biotechnol J* 6:2–12
 35. Smyth S, McHughen A (2008) Regulating innovative crops technologies in Canada: the case of regulation genetically modified crops. *Plant Biotechnol J* 6:213–225
 36. COGEM (2006) New techniques in plant biotechnology. COGEM Report CGM/061024-02. <http://www.cogem.net/ContentFiles/CGM061024-02%20new%20techniques%20in%20plantbiotechnology.pdf>
 37. Morris SH, Spillane C (2008) GM directive deficiencies in the European Union. *EMBO Rep* 9:500–504
 38. Batista R, Saibo N, Lourenço T, Oliveira MM (2008) Microarray analyses reveal that plant mutagenesis may induce more transcriptomic changes than transgene insertion. *Proc Natl Acad Sci USA* 105:3640–3645
 39. Kok EJ, Keijer J, Kleter GA, Kuiper HA (2008) Comparative safety assessment of plant-derived foods. *Regul Toxicol Pharmacol* 50:98–113
 40. Bradford KJ, Van Deynze A, Gutterson N, Parrott W, Strauss SH (2005) Regulating transgenic crops sensibly: lessons from plant breeding, biotechnology and genomics. *Nat Biotechnol* 23:439–444
 41. Morris SH (2007) EU biotech crop regulations and environmental risk: a case of the emperor's new clothes? *Trends Biotechnol* 25:2–6
 42. Jacobsen E, Schouten HJ (2007) Cisgenesis strongly improves introgression breeding and induced translocation breeding of plants. *Trends Biotechnol* 25:219–223
 43. Rommens CM, Haring MA, Swords K, Davies HV, Belknap WR (2007) The intragenic approach as a new extension to traditional plant breeding. *Trends Plant Sci* 12:397–403
 44. COGEM (2009) Novel plant breeding techniques – consequences of new genetic modification-based plant breeding techniques in comparison to conventional plant breeding, report 2009-02. <http://www.cogem.net/ContentFiles/Microsoft%20Word%20-%20New%20plant%20breeding%20techniques%20definitieve%20versie.pdf>
 45. Foster M, Berry P, Hogan J (2003) Market access issues for GM products. Australian Bureau of Agricultural and Resource Economics (ABARE) eReport 03.13, Canberra, Australia. <http://abareonlineshop.com/PdfFiles/PC12559.pdf>
 46. Kuiper HA, Kleter GA, Noteborn HPJM, Kok EJ (2002) Substantial equivalence – an appropriate paradigm for the safety assessment of genetically modified food. *Toxicology* 181–182:427–431
 47. Kok EJ, Kuiper HA (2003) Comparative safety assessment for biotech crops. *Trends Biotechnol* 21:439–444
 48. König A, Cockburn A, Crevel RWR, Debruyne E, Grafstroem R, Hammerling U, Kimber I, Knudsen I, Kuiper HA, Peijnenburg AACM, Penninks AH, Poulsen M, Schauzu M, Wal JM (2004) Assessment of the safety of foods derived from genetically modified (GM) crops. *Food Chem Toxicol* 42:1047–1088
 49. Paoletti C, Flamm E, Yan W, Meek S, Renckens S, Fellous M, Kuiper H (2008) GMO risk assessment around the world: some examples. *Trends Food Sci Technol* 19:S66–S74
 50. Devos Y, Lheureux K, Schiemann J (2010) Regulatory oversight and safety assessment of plants with novel traits. In: Kempken F, Jung C (eds) Genetic modification of plants – agriculture, horticulture & forestry, Series: Biotechnology in Agriculture and Forestry, vol 64. Springer, Berlin, pp 553–574
 51. Demeke T, Perry DJ, Scowcroft WR (2006) Adventitious presence of GMOs: scientific overview for Canadian grains. *Can J Plant Sci* 86:1–23
 52. Lheureux K, Mestdagh S, Devos Y (2008) Collaboration between the EFSA GMO Panel and national EU competent authorities for the environmental risk assessment of GM crop cultivation applications. *J Consum Prot Food Saf* 3(52):51
 53. SCP (2001) Opinion of the Scientific Committee on Plants concerning the adventitious presence of GM seeds in conventional seeds. http://ec.europa.eu/comm/food/fs/sc/scp/out93_gmo_en.pdf
 54. Verhoog H, Matze M, Lammerts Van Bueren E, Baars T (2003) The role of the concept of the natural (naturalness) in organic farming. *J Agric Environ Ethics* 16:29–49
 55. Altieri MA (2005) The myth of coexistence: why transgenic crops are not compatible with agroecologically based systems of production. *Bull Sci Technol Soc* 25:1–11
 56. Sanvido O, Romeis J, Bigler F (2007) Ecological impacts of genetically modified crops: ten years of field research and commercial cultivation. *Adv Biochem Eng Biotechnol* 107:235–278
 57. Schiemann J (2003) Co-existence of genetically modified crops with conventional and organic farming. *Environ Biosaf Res* 2:213–217
 58. Devos Y, Demont M, Sanvido O (2008) Coexistence in the EU–return of the moratorium on GM crops? *Nat Biotechnol* 26:1223–1225
 59. Devos Y, Demont M, Dillen K, Reheul D, Kaiser M, Sanvido O (2009) Coexistence of genetically modified (GM) and non-GM crops in the European Union. *Agron Sustain Dev* 29:11–30

60. EC (2003) Commission Recommendation of 23 July 2003 on guidelines for the development of national strategies and best practices to ensure the coexistence of genetically modified crops with conventional and organic farming. Off J Eur Communities L189:36–47
61. Friesen LF, Nelson AG, Van Acker RC (2003) Evidence of contamination of pedigreed canola (*Brassica napus*) seedlots in western Canada with genetically modified herbicide resistance traits. *Agron J* 95:1342–1347
62. Jørgensen T, Hauser TP, Jørgensen RB (2007) Adventitious presence of other varieties in oilseed rape (*Brassica napus*) from seed banks and certified seed. *Seed Sci Res* 17:115–125
63. Devos Y, Reheul D, De Schrijver A (2005) The co-existence between transgenic and non-transgenic maize in the European Union: a focus on pollen flow and cross-fertilization. *Environ Biosaf Res* 4:71–87
64. Hüsken A, Dietz-Pfeilstetter A (2007) Pollen-mediated intra-specific gene flow from herbicide resistant oilseed rape (*Brassica napus* L.). *Transgenic Res* 16:557–569
65. Sanvido O, Widmer F, Winzeler M, Streit B, Szerencsits E, Bigler F (2008) Definition and feasibility of isolation distances for transgenic maize. *Transgenic Res* 17:317–355
66. Devos Y, Reheul D, De Schrijver A, Cors F, Moens W (2004) Management of transgenic herbicide-tolerant oilseed rape in Europe: a case study on minimizing vertical gene flow. *Environ Biosaf Res* 3:135–148
67. Lutman PJW, Berry K, Payne RW, Simpson E, Sweet JB, Champion GT, May MJ, Wightman P, Walker K, Lainsbury M (2005) Persistence of seeds from crops of conventional and herbicide tolerant oilseed rape (*Brassica napus*). *Proc R Soc Lond B* 272:1909–1915
68. Messéan A, Sausse C, Gasquez J, Darmency H (2007) Occurrence of genetically modified oilseed rape seeds in the harvests of subsequent conventional oilseed rape over time. *Eur J Agron* 27:115–122
69. Gruber S, Colbach N, Barbottin A, Pekrun C (2008) Post-harvest gene escape and approaches for minimizing it. *CAB Rev Perspect Agric Vet Sci Nutr Nat Resour* 3:1–17
70. Jørgensen RB (2007) Oilseed rape: co-existence and gene flow from wild species. *Adv Bot Res* 45:451–464
71. Knispel AL, McLachlan S, Van Acker R, Friesen LF (2008) Gene flow and multiple herbicide resistance in escaped canola populations. *Weed Sci* 56:72–80
72. Pivard S, Adamczyk K, Lecomte J, Lavigne C, Bouvier A, Deville A, Gouyon PH, Huet S (2008) Where do the feral oilseed rape populations come from? A large-scale study of their possible origin in a farmland area. *J Appl Ecol* 45:476–485
73. Beckmann V, Soregaroli C, Wesseler J (2006) Coexistence rules and regulations in the European Union. *Am J Agric Econ* 88:1193–1199
74. EC (2006) Report on the implementation of national measures on the co-existence of genetically modified crops with conventional and organic farming, European Commission. http://ec.europa.eu/agriculture/coexistence/sec313_en.pdf
75. Koch BA (2007) Liability and compensation schemes for damage resulting from the presence of genetically modified organisms in non-GM crops. http://ec.europa.eu/agriculture/analysis/external/liability_gmo/full_text_en.pdf
76. Sundstrom F, Williams J, Van Deynze A, Bradford KJ (2002) Identity preservation of agricultural commodities. ANR Publication No. 8077, University of California
77. Czarnak-Klos M, Rodríguez-Cerezo E (2010) Best practice documents for coexistence of genetically modified crops with conventional and organic farming. 1. Maize crop production. European Coexistence Bureau (ECob) report. <http://ecob.jrc.ec.europa.eu/documents/Maize.pdf>
78. Demont M, Devos Y (2008) Regulating coexistence of GM and non-GM crops without jeopardizing economic incentives. *Trends Biotechnol* 26:353–358
79. Demont M, Daems W, Dillen K, Mathijs E, Sausse C, Tollens E (2008) Regulating coexistence in Europe: Beware of the domino-effect! *Ecol Econ* 64:683–689
80. Demont M, Dillen K, Daems W, Sausse C, Tollens E, Mathijs E (2009) On the proportionality of EU spatial *ex ante* coexistence regulations. *Food Pol* 34:508–518
81. Demont M, Devos Y, Sanvido O (2010) Towards flexible coexistence regulations for GM crops in the EU. *EuroChoices* 9:18–24
82. Demont M, Dillen K, Daems W, Sausse C, Tollens E, Mathijs E (2010) On the proportionality of EU spatial *ex ante* coexistence regulations: Reply. *Food Pol* 35:183–184
83. Beckmann V, Soregaroli C, Wesseler J (2010) *Ex-ante* regulation and *ex-post* liability under uncertainty and irreversibility: governing the coexistence of GM crops. *Econ J* 4, <http://www.economics-ejournal.org/economics/journalarticles/2010-9>
84. Ramessar K, Capell T, Twyman RM, Christou P (2010) Going to ridiculous lengths: European coexistence regulations for GM crops. *Nat Biotechnol* 28:133–136
85. Riesgo L, Areal FJ, Sanvido O, Rodríguez-Cerezo E (2010) Distances needed to limit cross-fertilization between GM and conventional maize in Europe. *Nat Biotechnol* 28:780–782
86. Craig W, Tepfer M (2007) Introduction to the safety/risk assessment of GM crops. *J Agric Invest* 5:36–42
87. EFSA (2006) Guidance document of the scientific panel on genetically modified organisms for the risk assessment of genetically modified plants and derived food and feed. *EFSA J* 99:1–100
88. Hill RA (2005) Conceptualizing risk assessment methodology for genetically modified organisms. *Environ Biosaf Res* 4:67–70
89. Conner AJ, Glare TR, Nap J-P (2003) The release of genetically modified crops into the environment. Part II: overview of ecological risk assessment. *Plant J* 33:19–46
90. Craig W, Tepfer M, Degrassi G, Ripandelli D (2008) An overview of general features of risk assessment of genetically modified crops. *Euphytica* 164:853–880
91. Hill RA, Sendashonga C (2003) General principles for risk assessment of living modified organisms: lessons from chemical risk assessment. *Environ Biosaf Res* 2:81–88

92. Wolt JD, Keese P, Raybould A, Fitzpatrick JW, Burachik M, Gray A, Olin SS, Schiemann J, Sears M, Wu F (2010) Problem formulation in the environmental risk assessment for genetically modified plants. *Transgenic Res* 19:425–436
93. EPA (1998) Guidelines for ecological risk assessment. EPA/630/R095/002F, US EPA risk assessment forum. US Environmental Protection Agency, Washington DC. <http://cfpub.epa.gov/ncea/raf/recordisplay.cfm?deid=12460>
94. Raybould A (2006) Problem formulation and hypothesis testing for environmental risk assessments of genetically modified crops. *Environ Biosaf Res* 5:119–125
95. Raybould A (2007) Ecological versus ecotoxicological methods for assessing the environmental risks of transgenic crops. *Plant Sci* 173:589–602
96. Romeis J, Hellmich RL, Candolfi MP, Carstens K, De Schrijver A, Gatehouse AMR, Herman RA, Huesing JE, McLean MA, Raybould A, Shelton AM, Waggoner A (2010) Recommendations for the design of laboratory studies on non-target arthropods for risk assessment of genetically engineered plants. *Transgenic Res*. doi:10.1007/s11248-010-9446-x
97. Garcia-Alonso M, Jacobs E, Raybould A, Nickson TE, Sowig P, Willekens H, Van der Kouwe P, Layton R, Amijee F, Fuentes AM, Tencalla F (2006) A tiered system for assessing the risk of genetically modified plants to non-target organisms. *Environ Biosaf Res* 5:57–65
98. Horak MJ, Rosenbaum EW, Woodrum CL, Martens AB, Mery RF, Cothren JT, Burns JA, Nickson TE, Pester TA, Jiang C, Hart JL, Sammons B (2007) Characterization of roundup ready flex cotton, 'MON 88913', for use in ecological risk assessment: Evaluation of seed germination, vegetative and reproductive growth, and ecological interactions. *Crop Sci* 47:268–277
99. Nickson TE (2008) Planning environmental risk assessment for genetically modified crops: problem formulation for stress-tolerant crops. *Plant Physiol* 147:494–502
100. Garcia-Alonso M (2010) Current challenges in environmental risk assessment: The assessment of unintended effects of GM crops on non-target organisms. *IOBC/wprs Bull* 52:57–63
101. Raybould A, Tuttle A, Shore S, Stone T (2010) Environmental risk assessment for transgenic crops producing output trait enzymes. *Transgenic Res* 19:595–609
102. Romeis J, Bartsch D, Bigler F, Candolfi MP, Gielkens MMC, Hartley SE, Hellmich RL, Huesing JE, Jepson PC, Layton R, Quemada H, Raybould A, Rose RI, Schiemann J, Sears MK, Shelton AM, Sweet J, Vaituzis Z, Wolt JD (2008) Assessment of risk of insect-resistant transgenic crops to nontarget arthropods. *Nat Biotechnol* 26:203–208
103. Suter GW (2000) Generic assessment endpoints are needed for ecological risk assessment. *Risk Anal* 20:173–178
104. Sanvido O, Romeis J, Bigler F (2009) An approach for post-market monitoring of potential environmental effects of Bt-maize expressing Cry1Ab on natural enemies. *J Appl Entomol* 133:236–248
105. Suter GW (2007) Ecological risk assessment, 2nd edn. CRC Press, Boca Raton, p 643
106. Perry JN, ter Braak CJF, Dixon PM, Duan JJ, Hails RS, Huesken A, Lavielle A, Marvier M, Scardi M, Schmidt K, Tothmeresz B, Schaarschmidt F, van der Voet H (2009) Statistical aspects of environmental risk assessment of GM plants for effects on non-target organisms. *Environ Biosaf Res* 8:65–78
107. Storkey J, Bohan DA, Haughton AJ, Champion GT, Perry JN, Poppy GM, Woiwod IP (2008) Providing the evidence base for environmental risk assessments of novel farm management practices. *Environ Sci Policy* 11:579–587
108. EFSA (2010) EFSA Panel on Plant Protection Products and their Residues (PPR); Scientific Opinion on the development of specific protection goal options for environmental risk assessment of pesticides, in particular in relation to the revision of the Guidance Documents on Aquatic and Terrestrial Ecotoxicology (SANCO/3268/2001 and SANCO/10329/2002). *EFSA J* 1821, 1–55. <http://www.efsa.europa.eu/en/scdocs/doc/1821.pdf>
109. Marvier M (2002) Improving risk assessment for nontarget safety of transgenic crops. *Ecol Appl* 12: 1119–1124
110. Lövei GL, Arpaia S (2005) The impact of transgenic plants on natural enemies: a critical review of laboratory studies. *Entomol Exp Appl* 114:1–14
111. Qi A, Perry JN, Pidgeon JD, Haylock LA, Brooks DR (2008) Cost-efficacy in measuring farmland biodiversity – lessons from the farm scale evaluations of genetically modified herbicide-tolerant crops. *Ann Appl Biol* 152: 93–101
112. Gathmann A, Wirooks L, Hothorn LA, Bartsch D, Schuphan I (2006) Impact of Bt-maize pollen (MON810) on lepidopteran larvae living on accompanying weeds. *Mol Ecol* 15: 2677–2685
113. Todd JH, Ramankutty P, Barraclough EI, Malone LA (2008) A screening method for prioritizing non-target invertebrates for improved biosafety testing of transgenic crops. *Environ Biosaf Res* 7:35–56
114. Aviron S, Sanvido O, Romeis J, Herzog F, Bigler F (2009) Case-specific monitoring of butterflies to determine potential effects of transgenic Bt-maize in Switzerland. *Agric Ecosyst Environ* 131:137–144
115. Rauschen R, Schultheis E, Pagel-Wieder S, Schuphan I, Eber S (2009) Impact of Bt-corn MON88017 in comparison to three conventional lines on *Trigonotylus caelestialium* (Kirkaldy) (Heteroptera: Miridae) field densities. *Transgenic Res* 18:203–214
116. Andow DA, Lövei GL, Arpaia S (2006) Ecological risk assessment for Bt crops. *Nat Biotechnol* 24:749–751
117. Raybould A (2010) Reducing uncertainty in regulatory decision-making for transgenic crops. More ecological research or clearer environmental risk assessment? *GM Crops* 1:1–7

118. EFSA (2010) Scientific opinion of the scientific panel on genetically modified organisms on statistical considerations for the safety evaluation of GMOs. EFSA J 1250:1–59
119. OECD (1993) Safety evaluation of foods derived by modern biotechnology, concepts and principles. Organization for Economic Co-operation and Development, Paris. ISBN 92-64-13859-5
120. Petersen W, Umbeck P, Hokanson K, Halsey M (2005) Biosafety considerations for selectable and scorable markers in cassava (*Manihot esculenta* Crantz) biotechnology. Environ Biosaf Res 4:89–102
121. Aumaitre A, Aulrich K, Chesson A, Flachowsky G, Piva G (2002) New feeds from genetically modified plants: substantial equivalence, nutritional equivalence, digestibility, and safety for animals and the food chain. Livest Prod Sci 74:223–238
122. Andow DA, Zwahlen C (2006) Assessing environmental risks of transgenic plants. Ecol Lett 9:196–214
123. Sanvido O, Widmer F, Winzeler M, Bigler F (2005) A conceptual framework for the design of environmental post-market monitoring of genetically modified plants. Environ Biosaf Res 4:13–27
124. Schmidt K, Wilhelm R, Schmidtke J, Beissner L, Mönkemeyer W, Böttinger P, Sweet J, Schiemann J (2008) Farm questionnaires for monitoring genetically modified crops: a case study using GM maize. Environ Biosaf Res 7:163–179
125. Wilhelm R, Sanvido O, Castanera P, Schmidt K, Schiemann J (2009) Monitoring the commercial cultivation of Bt maize in Europe – conclusions and recommendations for future monitoring practice. Environ Biosaf Res 8:219–225
126. Sweet JB (2008) General surveillance of multiple transgenic events. J Verbr Lebensm 3:12
127. Graef F, De Schrijver A, Murray A (2008) GMO monitoring data coordination and harmonisation at EU level – outcomes of the European Commission Working Group on Guidance Notes supplementing Annex VII of Directive 2001/18/EC. J Verbr Lebensm 3:17–20
128. ILSI Task Force (2004) Nutritional and safety assessments of foods and feeds nutritionally improved through biotechnology. Task Force of the International Life Science Institute (ILSI) International Food Biotechnology Committee. Compr Rev Food Sci Food Saf 3:35–104
129. Dutton A, Romeis J, Bigler F (2003) Assessing the risks of insect resistant transgenic plants on entomophagous arthropods: Bt-maize expressing Cry1Ab as a case study. BioControl 48:611–636
130. Wilkinson MJ, Sweet J, Poppy GM (2003) Risk assessment of GM plants: avoiding gridlock? Trends Plant Sci 8:208–212
131. Bartsch D, Gathmann A, Saeglitz C, Sinha A (2008) Field testing of transgenic plants. In: Stewart CN Jr (ed) Plant biotechnology and genetic principles, techniques, and applications. Wiley, London, pp 311–323
132. ACRE (2007) Managing the footprint of agriculture: towards a comparative assessment of risks and benefits for novel agricultural systems. DEFRA, London. <http://www.defra.gov.uk/environment/acre/fsewiderissues/pdf/acre-wi-final.pdf>
133. EFSA (2008) Environmental risk assessment of genetically modified plants – challenges and approaches. EFSA Scientific Colloquium Series 8, June 2007. European Food Safety Authority, Brussels. <http://www.efsa.europa.eu/en/colloquiareports/colloquiagmoera.htm>
134. Sanvido O, Romeis J, Gathmann A, Gielkens M, Raybould A, Bigler F (2011) Evaluating environmental risks of genetically modified crops – ecological harm criteria for regulatory decision-making. Environment & Science Policy, doi:10.1016/j.envsci.2011.08.006

Transgenic Fishes: Applications, State of the Art, and Risk Concerns

ERIC M. HALLERMAN

Department of Fisheries and Wildlife Sciences, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

Article Outline

Glossary

Definition of the Subject

Introduction

Ecological Risk Assessment for Transgenic Fish in the Accessible Ecosystem

Risk Management

Future Directions: Transgenic Fish Considered Within the Sustainability Paradigm

Bibliography

Glossary

Accessible ecosystem The aquatic environment immediately accessible to an organism were it to escape from the research or culture facility, and more distant habitats in the contiguous environment into which the organism or its offspring reasonably may be expected to disperse.

Confinement A system of physical, chemical, biological, and operations management measures used to contain a transgenic fish within an experimental or culture facility or to prevent its establishment in the accessible ecosystem upon escape or release.

Direct effects (or impacts) Effects of a transgenic fish on an (other) organism(s) in the accessible

ecosystem, which are effected through mechanisms involving biotic factors. Examples would include but not be limited to reproduction with or predation upon or competition with members of the same or other species.

Environmental effects (or impacts) Consequences of an experiment or of aquaculture that might include, but are not limited to, changes in the structure, function, or resiliency of the accessible ecosystem; changes in the gene pool of populations resident in the accessible ecosystem; or decline in the abundance of a population of threatened, endangered, or special concern species in the accessible ecosystem.

Harm A perturbation resulting in negative impact to a population or species.

Hazard An agent or process that has the potential to produce harm.

Indirect effects (or impacts) Effects of a transgenic fish on (an) other organism(s) in the accessible ecosystem that are effected through mechanisms involving abiotic factors or additional species. Examples would include, but not be limited to, modification of the physical environment, affecting its suitability as habitat for other species and cascading effects of altered trophic function upon the aquatic community.

Introgression The incorporation of genes from one species into the gene pool of another; alternatively, breeding of a transgenic individual with wild-type individuals, leading to introduction and persistence of the transgene in the wild gene pool.

Novel trait Expression of a compound not normally found in the species, e.g., an antifreeze polypeptide in Atlantic salmon; alternatively, expression of a compound normally found in the species, but under novel gene regulation; for example, expression of a growth hormone gene under the transcriptional control of a promoter element not normally associated with the gene.

Risk The likelihood of harm resulting from exposure to a hazard.

Transgene An artificial DNA molecule containing a structural gene whose expression would confer a novel trait upon its host. The transgene would contain a regulatory element to control its expression in the host and an element signaling

the point at which transcription would terminate and a poly-A tail added to the messenger RNA transcript.

Definition of the Subject

A transgenic fish or shellfish bears within its chromosomal DNA a gene construct – that is, a transgene, a gene whose expression is under novel regulation – that was introduced by human intervention. A variety of methods may be utilized to introduce foreign DNA into the genome of fishes [1]. Microinjection of a DNA vector into the cytoplasm of a fertilized egg was the first [2, 3] and remains the most commonly used method for producing founders of a transgenic line. Electroporation, the application of short electrical pulses to make the cell membrane permeable and thereby permit entry of DNA vectors into fertilized fish eggs, has been used to produce transgenic fishes [4, 5]. Electroporation also has been used to introduce DNA vectors into sperm, which in turn was utilized to fertilize eggs to produce transgenic founders [6]. A variation of the “gene gun” method, which involves bombardment of fertilized eggs with DNA-coated microprojectiles, was demonstrated using medaka, or rice fish, a model species [7]. Retroviruses [8] and transposable elements [9] have been modified to carry genes of interest and, using the ability of these elements to insert themselves into the genome of a host, to insert foreign DNA into zebrafish chromosomes.

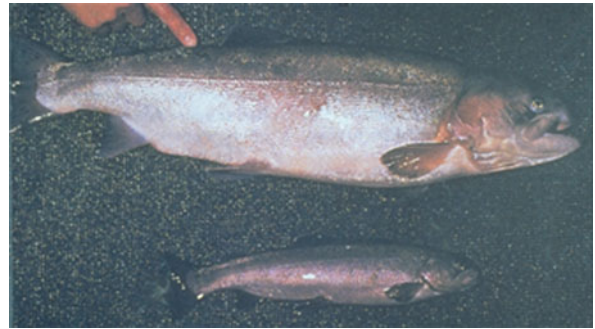
Because they have large eggs, high fecundity, and external fertilization and development, fishes provide excellent systems for gene transfer. By introducing a major gene whose expression is under novel regulation, biotechnologists can achieve dramatic impacts upon valued phenotypes. Since the mid-1980s when the first transgenic fish were produced [2, 3], gene transfer studies have been conducted using over 35 species [1]. Following the demonstration of the concept [10], a variety of transgenic zebrafish, medaka, and other fishes have been developed as experimental models for biomedical research [11–13]. Many transgenic lines have been developed for potential use in aquaculture (see below), targeting growth rate enhancement and a variety of other production-related traits. Similarly, several sport fishes (e.g., northern pike [14]) have been transformed with growth hormone

genes in order to increase growth rate, aggressiveness, or ultimate size. Transgenic zebrafish expressing fluorescent proteins have been marketed as novel aquarium pets [15]. Transgenic lines have been developed for biomonitoring of chemicals in the aquatic environment [16–18].

Because they express novel traits, transgenic fish may prove attractive for aquaculture, fisheries management, or other applications. While production of transgenic fishes may prove profitable, there are broader and longer-term questions regarding whether such use would prove ecologically, economically, and socially sustainable. Because most transgenic fish lines have been produced for purposes of aquaculture, this contribution focuses upon sustainability of transgenic fishes intended for food production.

Introduction

As world population increases, so also does its demand for fisheries products. However, since 1985, the total harvest from capture fisheries has fluctuated at approximately 90 million metric tons (MMT) [19], suggesting that capture fisheries are at or above maximum sustainable yields [20]. Since 1970, the contribution of aquaculture to world supplies of fisheries products has grown from approximately 5 MMT to over 50 MMT annually. Against this background, it is clear that the projected growth in demand for fisheries products can be met only through growth in production from aquaculture. The historical and projected growth in aquaculture raises the question of whether aquaculture is sustainable as currently practiced. First raised by the environmental community [21], the question then was critically examined by the aquaculture community itself [22]. Academic and commercial aquaculturists now are working to define and achieve sustainability [23–25]. Among key innovations so driven are changes in aquaculture feeds to decrease the content of scarce fish meal and oil [26, 27] and selective breeding to develop stocks of fish adapted to utilizing such feeds [28, 29]. There is increasing emphasis among aquaculturists on producing fishes that are herbivores or omnivores rather than carnivores. Within this broader context, whether transgenic fishes would contribute to the sustainability of aquaculture depends upon the host fish, the transgene, the confinement system in



Transgenic Fishes: Applications, State of the Art, and Risk Concerns. Figure 1

A coho salmon expressing a salmon growth hormone transgene (*above*) exhibited growth that was dramatically greater than that of a same-aged, non-transgenic full-sibling (*below*) (Courtesy: R.H. Devlin)

which the fish are raised, and any impacts that might follow should transgenic fish escape or be introduced into the accessible environment.

Traits targeted by transgenesis. A number of different traits have been targeted for genetic improvement via transgenesis. Growth hormone (GH) genes have been introduced into over a dozen aquaculture species to increase growth rate. Noteworthy examples where growth was increased several-fold relative to wild-type include Atlantic salmon [30, 31], coho salmon [32, Fig. 1], Nile tilapia [33], common carp [34], and rohu carp [35]. The most notable example of growth enhancement is that of mud loach, for which up to 35-fold growth rate enhancement was observed [36]. Antifreeze protein genes from Arctic or Antarctic fishes have been transferred to increase the tolerance of species sensitive to freezing in marine waters chilled to below the freezing point of fish flesh (e.g., Atlantic salmon [37]). Genes for bactericidal compounds have been introduced to heighten nonspecific disease resistance of cultured fishes (e.g., cecropin in channel catfish [38] and human lactoferrin in grass carp [39, 40]). The major storage form for phosphorus in grain is phytate, which is largely indigestible and has anti-nutritional properties; following demonstration that medaka transgenic for a microbial phytase gene exhibited improved phytate utilization [41], phytase also was transferred into tilapia [42]. Transgenic

approaches have been proposed for reproductive confinement of fish [43] (see also below), including expression of an antisense DNA complementary to the gonadotropin-releasing hormone transcript [44] and a “sterile-feral” *zBMP2* construct that blocks a critical developmental gene unless a repressor compound is applied to the animal in captivity [45]. When a key biochemical pathway is disrupted by lack of one enzyme, the function of that pathway can be restored by the expression of a transgene; for example, a rat L-gulonono- γ -lactone oxidase transgene was introduced into rainbow trout in an attempt to complete the biosynthetic pathway for L-ascorbic acid [46]. While not exhaustive, these examples illustrate the scope of efforts to genetically improve cultured stocks of fishes using gene transfer methodologies.

Most transgenic lines described above have not been subject to the generations of breeding needed to develop a homozygous line stably expressing the transgene. However, development of some GH-transgenic lines is well advanced and efforts to commercialize them are ongoing. A/F protein, producers of GH-transgenic Atlantic salmon, have petitioned the US Food and Drug Administration for approval to market eggs of transgenic Atlantic salmon to commercial aquaculture producers [31]. The Cuban government is considering a request for production of GH-transgenic tilapia [47]. Marketing of a GH-transgenic common carp line is under consideration by the Chinese government [48]. With the prospect of improved production efficiency, it is not surprising that some aquaculturists want to produce GH-transgenic fish commercially. Commercialization of transgenic fishes, however, poses ecological, food safety, regulatory, and animal welfare concerns [49–52].

Confinement systems. Many different fish culture systems are in commercial use. Closed recirculating systems hold fish in indoor tanks and the water is treated in filtration systems; the associated expenses are such that only high-value species can be produced economically. Raceways are long, narrow outdoor tanks, with water from a spring routed through the system once and then discharged; rainbow trout are commonly reared in raceways. Ponds, which may have discharge to a nearby stream or river, are used to produce many species including tilapias and catfishes. Marine netpens, mesh cages suspended in bays or

fjords, are used to produce Atlantic salmon, coho salmon, Atlantic cod, and other species. Ocean ranching involves release of juvenile salmon, trout, or charr into the ocean, to be collected later upon their return to spawn as adults. Clearly, these different types of culture systems vary in terms of their production economics and in their ability to reliably confine fishes.

Entry of cultured fish into the accessible ecosystem. Most commercial aquaculture operations have a routine, often significant escape of fish. Escapes can occur through equipment failures, during handling or transport operations, through predator intrusion into facilities, as a result of storms, or by other mechanisms. In particular, escapes of salmon and trout from marine net-pen facilities are common, ranging from minor incidents where a few fish escape to massive escapes. In the native range of Atlantic salmon, an estimated two million farmed salmon escape each year into the North Atlantic [53]. Outside the native range, millions of Atlantic salmon have escaped on the west coasts of North [54] and South America [55]. Although salmon farm operators are attempting to prevent escapes by upgrading confinement systems, installing predator deterrent devices, and other actions, it still must be assumed that escapes will occur [56].

Escape of cultured fish into the accessible ecosystem and interaction with local intraspecific and interspecific populations is a well-documented environmental concern [57–61]. The potential ecological and genetic interactions between cultured and wild fish have raised concern about the viability of affected wild stocks. Ecological concerns focus upon competition for space and food resources and direct predation [62]. Genetic concerns include the potential breakdown of locally adapted traits through interbreeding and introgression and range up to replacement of native stocks by cultured stocks [63]. Such concerns are posed by the prospect of producing transgenic fish in aquaculture, with the additional unknown posed by possible effects of the transgene. Because of the prospect of escape of cultured transgenic fish, we must consider not only the benefits of transgenic fish to aquaculture, but also any ecological and genetic hazards that they pose. Hazards posed by transgenic fish were first inferred on the basis of ecological principles [64–67], and then supported by empirical studies.

Ecological Risk Assessment for Transgenic Fish in the Accessible Ecosystem

Ecological risk assessment for transgenic organisms should be based upon case-by-case assessment of the host species, transgene construct, transgene integration site within the genome, and receiving ecosystem [49, 64, 68, 69]. Because most fishes do not have a long history of domestication, many cultured stocks retain the ability to survive and reproduce in the accessible ecosystem. Hence, consideration of ecological risk and genetic hazards posed by transgenic fish that might escape from a culture facility is appropriate.

Consideration of harms posed by transgenic fish must be based on an understanding of key concepts underlying the science and practice of risk analysis [49]. In this context, a *harm* is defined as a perturbation resulting in negative impacts to a species. A *hazard* is defined as an *agent or process* that has the potential to produce harm. A *risk* is defined as the *likelihood* of harm resulting from exposure to the hazard. Risk, R , is estimated as the product of the probability of exposure, $P(E)$, and the conditional probability of harm given that exposure has occurred, $P(H|E)$. That is, $R = P(E) \times P(H|E)$. The steps in risk analysis, then, are to: (1) identify potential harms; (2) identify hazards that might lead to harms; (3) define what exposure means for an aquaculture stock and assess the likelihood of exposure, $P(E)$; (4) quantify the likelihood of harm given that exposure has occurred, $P(H|E)$; and (5) multiply the resulting probabilities to yield a quantitative estimate of risk.

Exact probabilities of risk may prove difficult or impossible to determine for all types of possible ecological or genetic harm. Indeed, it is unlikely that all possible harms would be known *a priori*, particularly with respect to any indirect effects. Hence, it may be necessary – based on current knowledge of population genetics, population dynamics, receiving ecological communities, and experience with cultured stocks – to classify levels of concern regarding likely ecological and genetic impacts posed by cultured stocks into *qualitative* categories ranging from low to high.

Potential harms. Harms potentially posed by transgenic fishes would be the outcome of a chain of events occurring after escape or release from a culture system. Examples of harms that potentially might be realized by

transgenic fishes entering or becoming established in an accessible ecosystem [49, 64–66] would include decline in abundance or loss of a native species. A locally adapted natural population could be replaced with a transgenic population or experience introgression of the transgene, decreasing the degree of local adaptation. Changes in ecosystem structure or function could result in decreased fisheries production or in ecosystem resiliency in the face of further ecological stressors.

Ecological hazards. Ecological hazards that could lead to harms becoming realized are posed to a range of species with which a transgenic fish might interact in the accessible ecosystem. Ecological hazards include the possibility of heightened predation or competition, as well as alteration of population or community dynamics due to activities of the transgenic fish. Examples from empirical studies illustrate some of these potential ecological hazards.

Several studies have focused on Atlantic salmon expressing a GH transgene. In order to support their rapid growth, these transgenic salmon have significantly faster routine metabolic rates and growth rates than domesticated and wild individuals of equal mass, and therefore require more energy to sustain body function [70, 71]. GH-transgenic salmon must consume food at a more rapid rate than control salmon. Abrahams and Sutterlin [72] observed behavior of size-matched transgenic and control salmon in experiments where fish could feed in safety or in the presence of the predator. Growth-enhanced transgenic fish were significantly more willing to risk exposure to a predator in order to gain access to food, and exhibited increased feeding rate and average speed of movement. However, transgenic fish reduced their exposure to predators in response to increases in the magnitude of risk, suggesting that their more active behavior may not necessarily lead to increased susceptibility to predation.

Devlin et al. [73] compared the intake of contested food pellets by size-matched pairs of one control (1 year older non-transgenic) and one transgenic coho salmon. The transgenic fish consumed 2.5 times more contested pellets than controls. Overall, transgenic fish consumed 2.9 times more pellets than non-transgenic controls, indicating high feeding motivation of transgenics throughout the feeding trials. In subsequent experiments, at low levels of food availability,

dominant, generally transgenic individuals dominated acquisition of food resources [74]. Further, when food availability was low, populations containing transgenics crashed, while in contrast, groups containing only non-transgenic salmon exhibited 72% survival and increase of population biomass. Hence, genotype-by-environment interactions affect ecological risk assessment. Sundstrom et al. [75] exposed GH transgenic and wild coho salmon fry to a live predator in a naturalized experimental stream tank under conditions of high and low food abundance. In both cases, mortality rates of transgenic fry were significantly higher than those of wild fry.

Ecological hazards posed by transgenic fish are not limited to salmonid species. Duan et al. [76] observed food consumption, frequency of movement, and feeding hierarchy in transgenic and size-matched control common carp under conditions of limited food supply. Transgenic fish consumed 1.9 times as many pellets, moved 73% more frequently, and exhibited a higher standing in the feeding hierarchy, but did not realize their higher growth potential. The authors concluded that elevated ability to compete for limited food resources could be advantageous to transgenic carp after escape into the accessible ecosystem.

These and other empirical findings to date tend to support the competitive ability of GH-transgenic fishes with conspecifics, and suggest the likelihood that ecological harms could become realized should large numbers of GH transgenics escape from culture or a transgenic population become established. Clearly, not all possible ecological interactions have been considered experimentally. Ecological hazards posed by transgenic fish expressing other sorts of transgenes, for example, a phytase gene that could confer ability to effectively utilize a broader range of foods, have not yet been examined empirically.

Given the range of interacting species and accessible ecosystems, and given the temporal and spatial variation of ecosystems, it is difficult to predict the ecological outcome should transgenic fish escape from aquaculture operations and enter accessible ecosystems. Devlin et al. [77] review methodologies to assess potential ecosystem effects of transgenic fish before they enter unconfined environments, outlining a strategy for identifying the most important data and discussing methods for obtaining

them. Key issues include characterization of the accessible ecosystem, phenotypic characterization of the transgenic fish, experimentation appropriate for relating the transgenic phenotype and ecological impacts, and recognizing and accounting for sources of uncertainty.

Genetic hazards. Although reproductive confinement methodologies such as sterilization by triploidy or transgenesis have been proposed to minimize reproductive interaction between transgenic and wild fish (see below), to date these technologies have not been demonstrated to be completely effective [78]. Should transgenic fish be reproductively fertile, they potentially could interbreed with natural populations. Genetic or evolutionary impacts that could be realized would depend on the fitness of novel genotypes in the wild. There may be cases where fitness relative to the wild type is high, posing introgression of transgenes into natural populations. There may also be cases where maladaptive traits would be introduced into native populations, posing a hazard to the demographic viability of the receiving population. Empirical studies illustrate possible genetic hazards to wild populations of fish.

Considering the possible impact of introgression of transgenes into wild populations, Devlin et al. [79] examined growth enhancement due to expression of a GH transgene in both wild and selectively bred commercial rainbow trout strains and showed that the effect of expression of the transgene varied among the respective strains. Transgenic wild-strain rainbow trout retained the slender body morphology of the wild-type strain, but their final size at maturity was increased by transgenesis. Both domestic and wild-strain trout had reduced viability, and in the case of the domestic strain, all transgenic individuals died before sexual maturation. The greatest response to expression of the transgene was in hybrids of a wild and domesticated strain.

Devlin et al. [80] compared the reproductive performance of transgenic and non-transgenic coho salmon. GH-transgenic fish showed precocious smoltification and onset of sexual maturation but no increase of adult body size, indicating compression of the normal life history. However, strong genotype-by-environment interactions were shown for the effects of transgenesis and culture environment. Bessey et al. [81]

found no differences in gamete quality or in vitro offspring production. The transgenic fish were able to breed successfully in the absence of competition, but at a significantly lower level than hatchery fish (i.e., fish reared artificially as young and then released to complete their life history naturally). When in direct competition with hatchery fish, however, the transgenic coho salmon were competitively and reproductively inferior to the point where they had little or no success.

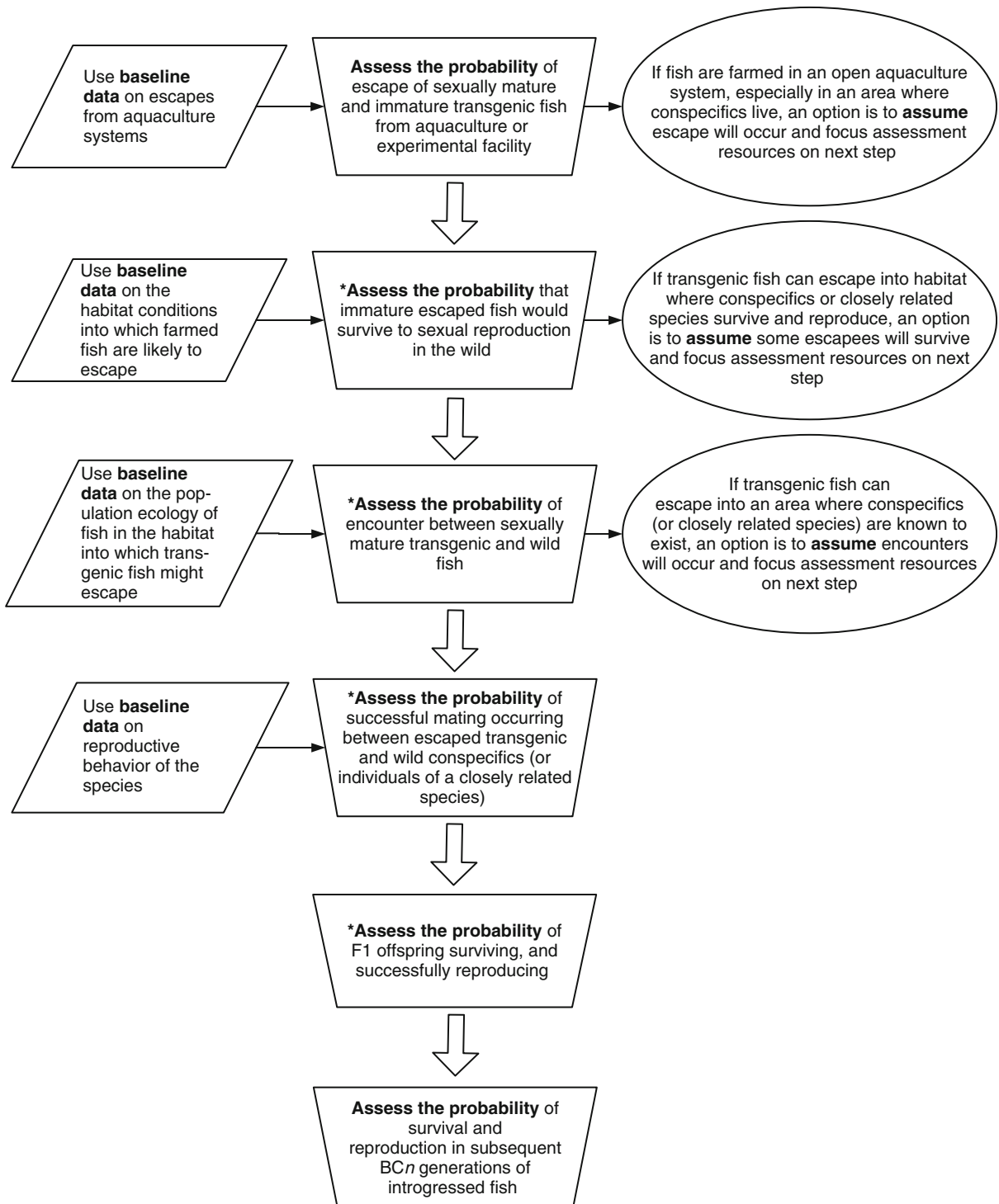
Models have been developed to predict the outcome of introgression of a transgene into a wild population, notably for cases where expression of the transgene differentially affects various components of fitness. Muir and Howard [82] developed a deterministic model that predicted that under certain conditions, a transgene introduced into a natural population by a small number of transgenic fish would spread as a result of enhanced mating advantage enjoyed by larger individuals, but the reduced viability of offspring would cause eventual local extinction of populations. The predicted time to extinction of a wild population would be a function of the mating advantage of transgenic males relative to that of wild-type males and the relative viability of transgenic offspring. Muir and Howard [83] subsequently considered a transgenic organism's fitness components (juvenile viability, adult viability, age at sexual maturity, female fecundity, male fertility, and mating advantage), using various combinations of fitness component values in addition to empirically derived estimates for medaka, a model species, to parameterize and run the model. For a wide range of parameter values, the model predicted that transgenes could spread through populations despite high juvenile viability costs if there were sufficiently high positive effects on other fitness components. Sensitivity analyses showed that transgene effects on age at sexual maturity would have the greatest effect on transgene frequency, followed by juvenile viability, mating advantage, female fecundity, and male fertility, with the least impact due to changes in adult viability. Extinction hazards also could be posed when the transgene: (1) increases male mating success but reduces adult viability; (2) increases adult viability but reduces male fertility; or (3) increases both male mating success and adult viability, but reduces male fertility [84]. Using a different approach, Hedrick [85] developed a deterministic model for invasion of transgenes into

natural populations and showed that if a transgene gives rise to a male mating advantage and a general viability disadvantage, then the conditions for its invasion of a natural population are very broad. More specifically, for two-thirds of the possible combinations of mating and viability parameters, the transgene increased in frequency. In addition, the demographic viability of the natural population was reduced, increasing the probability of extinction of the natural population.

Kapuscinski et al. [86] presented a step-by-step method for assessing risk posed by gene flow and possible introgression of a transgene into a wild population (Fig. 2) considering probabilities of escape of a transgenic fish from confinement, encountering wild-type mates, reproducing, and the young surviving to transmit the transgene to future generations. Currently, our empirical knowledge of genetic hazards posed by production of transgenic fish and shellfish and their associated risks is limited. Ecological risk assessments are ongoing for GH-transgenic coho and Atlantic salmon, as well as other transgenic fishes, including both model species and aquaculture species. Many critical experiments aimed at estimating fitness parameters [67, 83] have yet to be conducted.

Risk Management

Many key inputs for a quantitative ecological risk assessment currently are unknown. Further, many key aspects of risk assessment are difficult or impossible to address, given the spatial and temporal variability of ecosystems and the adaptive ability of populations. Under at least some circumstances, escaped transgenics could negatively impact accessible ecosystems and populations; hence, should a producer or oversight body determine that production of a transgenic stock poses harm to a population in the accessible ecosystem, the consideration then turns to managing the associated risk. From the viewpoint of ecological sustainability, risk might be managed by producing transgenic fish only under conditions of confinement. In some contexts, production of transgenic fish might go forward only under conditions of strict confinement aimed at ensuring no escape of transgenic fish into the accessible ecosystem.



Transgenic Fishes: Applications, State of the Art, and Risk Concerns. Figure 2

A pathway for conducting an assessment of the risk of gene flow posed by a transgenic fish (From [86]). Asterisks denote assessment steps that require empirical information on traits exhibited by the transgenic fish

Risk management is the design, selection, and implementation of a program of actions to minimize risk. In the context of formal risk analysis, it becomes clear that the best approach for minimizing the likelihood of harm being realized is to minimize exposure to the hazard. Three non-mutually exclusive approaches include: (1) physically confining the cultured stock on aquaculture facilities, (2) reproductively confining cultured stocks, and (3) operations management. Management of risks posed by transgenic fishes has been considered for experimental [87] and commercial production systems [88].

Physical confinement. Physical confinement of cultured aquatic organisms will require a combination of measures in order to prove effective [87, 88]. Context is key; the ease or difficulty of managing risk will depend greatly on the geographic location of an aquaculture facility. Sites subject to flooding, violent storms, or wave action are poorly suited for confinement of production stocks. Virtually all physical confinement systems will include barriers to escape of cultured organisms from the culture site, including mechanical or physical/chemical barriers. Mechanical barriers are structures that physically hold back cultured organisms from escaping the project site. Examples include stationary or moving screens, (e.g., floor drains, standpipe screens), tank covers, filters (e.g., gravel traps), grinders or pumps, and French drains. A French drain is a filter for screening effluent from an aquaculture facility that contains gravel and geotextiles through which even small life-stages cannot pass. Physical or chemical barriers use manipulation of physical (e.g., elevation of temperature) or chemical (e.g., administration of chlorine) attributes of effluent water to induce 100% mortality of any escaped organisms before they can reach the accessible ecosystem. The set of barriers must prevent escape of the hardest-to-retain life-stage held at the aquaculture operation, usually the smallest life-stage. Because no barrier is 100% effective at all times, for effective physical confinement, each possible escape path from the aquaculture facility would have redundant barriers to escape of cultured organisms. Barriers also must prevent access of predators that can carry cultured organisms off-site (e.g., avian predators) or damage ponds (e.g., muskrats), allowing escape of cultured organisms.

Reproductive confinement. A key element of many risk management strategies is reproductive confinement,

especially for cases where physical confinement alone is unlikely to prove effective. Two approaches, culture of monosex or sterile stocks, might be applied singly or in combination.

Procedures have been developed for producing monosex stocks of fish through direct, hormone-mediated sex reversal and indirectly by sex reversal, followed by progeny testing and selection of broodstock that yield monosex offspring. For example, production of all-female rainbow trout [89] and all-male Nile tilapia [90] stocks comprise growing segments of the respective aquaculture sectors. Production of monosex stocks may provide an acceptable high level of reproductive confinement in contexts where transgenic escapees have no possibility of encountering prospective mates in the accessible ecosystem. Production of monosex stocks will not provide reliable reproductive confinement if there is a population of the same species or of a closely related species with which escaped transgenics may hybridize in the accessible ecosystem. Further, for monosex production to provide reliable confinement, sex ratios would have to be absolutely monosex, an outcome which has not generally been achieved in commercial production [88].

Reproductive confinement also might be achieved through chromosome set manipulation. Fishes can be produced that have three, instead of the usual two chromosome sets. Such fish are said to be triploid, and are reproductively sterile because they produce gametes with unusual numbers of chromosomes. Should their eggs or sperm give rise to an embryo, the unmatched chromosomes would disrupt normal embryogenesis, leading to death of the embryo. All-triploid stocks can be produced most reliably by the crossing of diploid and tetraploid broodstock, although lack of tetraploid broodstock precludes the approach for many species. Alternatively, triploid stocks can be produced by *de novo* induction. *De novo* triploidy induction is not always 100% effective and, hence, triploid broods will have to be screened to determine whether they are indeed all-triploid [78]. This extra handling and screening add to the cost of seed-stock production.

Other approaches for reproductive confinement may become available in the future, including the possibility of reversible sterility through transgenesis. A transgene that produces antisense gonadotropin-releasing hormone

appears to cause sterility in male transgenic rainbow trout [44], but does not seem fully effective in females [91]. Use of a similar approach in tilapia sometimes greatly reduced fertility, but at other times was ineffective in both sexes (N. Maclean, quoted in [45]). Thresher et al. [45] developed a transgene that renders the host sterile, allowing normal development only if a repressor compound is applied during a particular life-history stage. The authors demonstrated its effectiveness in founder-generation transgenic zebrafish, channel catfish, and Pacific oyster. Definitive demonstration of this “sterile-fertile” approach would require successful trials in homozygous transgenic lines and demonstration of the effectiveness of the repressor under commercial aquaculture conditions. Wang and Van Eenennaam [43] reviewed developing transgenic approaches to reproductive confinement of transgenic fishes, in addition to the approaches above also considering gonad-specific transgene excision. They suggested that slow progress in development of transgenic reproductive confinement systems is attributable to limited fundamental knowledge of possible gene targets to inactivate in order to induce sterility, preclude gonadal development, or disrupt embryogenesis, as well as difficulties in implementing such highly technical approaches *in vivo*.

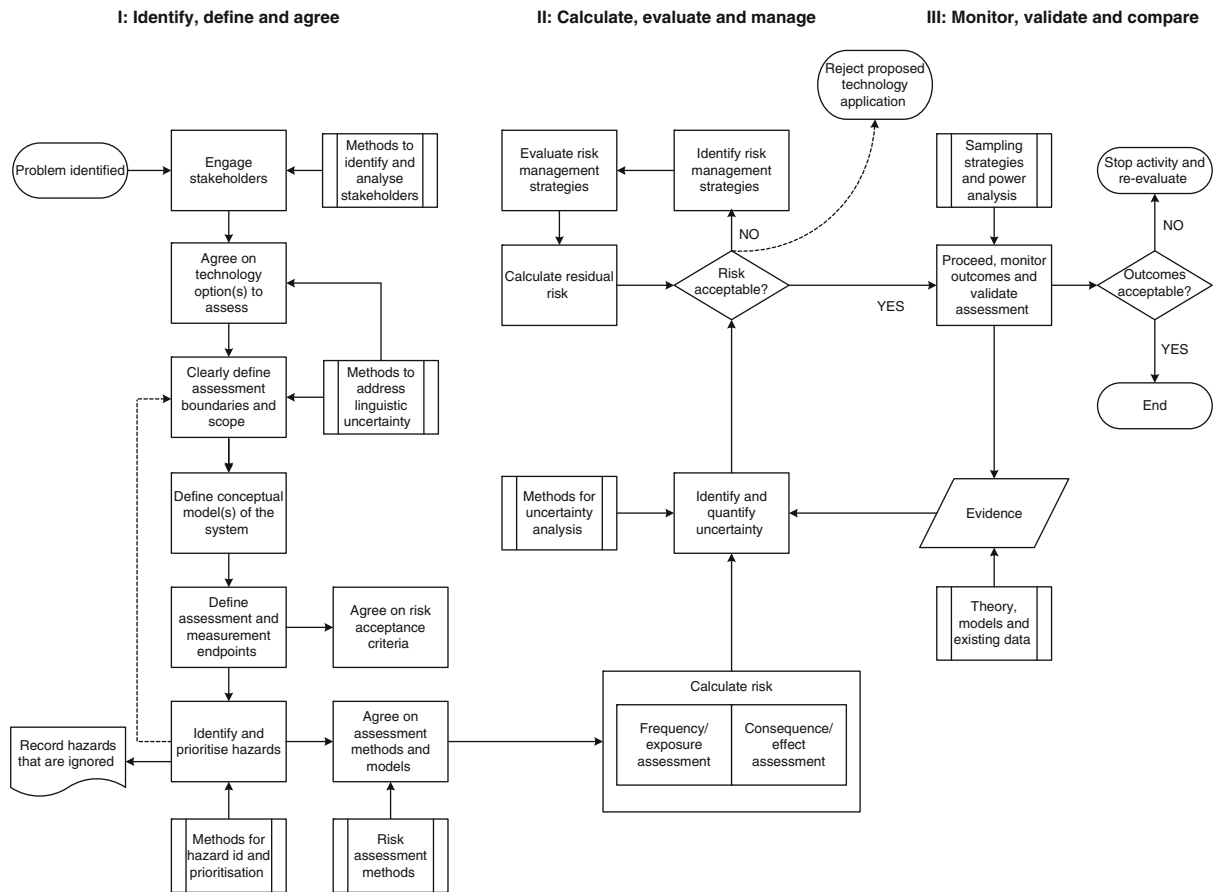
Operations management. Operations management is a key, though often overlooked, aspect of a confinement system [87, 88]. Measures are needed to: (1) ensure that normal activities of workers at the aquaculture operation are consistent with the goal of effective confinement, (2) prevent unauthorized human access to the site, and (3) ensure regular inspection and maintenance of physical confinement systems. Effective supervision of project personnel is critical for operations management.

Operations management also must consider biosecurity after cultured organisms are removed purposefully from the culture site, that is, harvested and transported through the marketing process. For biosecurity purposes, it would be best if only dead fish were sent to market. This is counter to marketing practices in many developing countries, where live sales prevail. Live sale is a known route for introductions of nonindigenous species, and exemplified by recent introduction and establishment of snakeheads in the USA [92].

Effective risk management calls for combinations of confinements. Combinations of risk management measures are advisable so that failure of any one measure will not necessarily lead to escape of confined stocks [87, 88]. It is infeasible to anticipate the best combination of risk management measures for every possible case. Differences in species, production traits, receiving ecosystems, and culture systems will affect the case-by-case determination of appropriate risk management measures.

Future Directions: Transgenic Fish Considered Within the Sustainability Paradigm

Interactive decision-making framework. Recognizing that sustainability has ecological, economic, and social dimensions, what conditions would have to be met for aquaculture production of transgenic fish to prove sustainable? The range of issues posed by a proposed utilization of transgenic fish in aquaculture might best be considered within a three-stage, interactive framework involving a range of stakeholders [93, 94] (Fig. 3). Involvement of the full range of stakeholders would bring all existing knowledge into the deliberative process, make the process transparent to stakeholders, enhance the understanding and acceptance of the outcome of risk analysis, and promote social acceptance of the technology. Stage I of the process involves identifying the problem at hand, engaging stakeholders, identifying possible technical solutions to the problem at hand, and identifying potential harms, risk pathways, and assessment methods. Stage II of the process is risk assessment, leading to estimating the likelihood that harm will become realized should the proposed production of transgenic fish go forward. Upon estimation of that risk, a decision would be faced as to whether the risk is acceptable. If it proves acceptable, the decision may be made to go forward. If the level of risk proves unacceptably high, risk management measures would be identified and residual risk quantified, and the decision of whether to go forward would again be considered. Should the proposed action be implemented, genetic, ecological, and social outcomes should be monitored. As discussed below, because all potential harms and associated pathways cannot be known and precisely predicted *a priori*, it will be necessary to update the risk analysis as knowledge accumulates using an adaptive management approach [95, 96].



Transgenic Fishes: Applications, State of the Art, and Risk Concerns. Figure 3

Overview of a framework for environmental risk assessment of a transgenic fish embedded within a larger social context (From [94]). The framework has three stages. In the first stage, participants agree of an assessment option, define the scope of the assessment, agree on conceptual models, identify assessment and measurement end points, and agree upon a list of prioritized hazards. In the second stage, the risks and uncertainties associated with these hazards are assessed, risks are compared to predetermined acceptance criteria, and where necessary, risk management strategies are identified and evaluated. In the third stage, predictions of the risk assessment are compared with empirical findings, generating data used to reexamine uncertainty in the risk assessment

Trade-offs of benefits and risks. The prospect of producing transgenic fishes in aquaculture poses a mix of benefits and risks that would affect the assessment of whether such production would promote ecological, economic, and social sustainability within the decision-making framework. Consideration of several examples will illustrate how trade-offs among benefits and risks will affect the assessment of sustainability.

Benefits and hazards posed by GH transgenics have been considered empirically, and some of the associated risks have been quantified. GH-transgenic fish

exhibit rapid growth and perhaps also improved feed conversion ratio, promoting production efficiencies in terms of more efficient utilization of inputs, including facilities, labor, and feedstuffs. However, as noted above, risk assessment research to date suggests a considerable likelihood of ecological or genetic impacts should GH transgenics escape into accessible ecosystems. This risk can be managed by adoption of an appropriate suite of risk management measures (see above). Risk management will have attendant costs, decreasing profitability by some as-yet unquantified

amount. Some segments of society are opposed to production of transgenic organisms [97], affecting the social aspect of whether this first wave of transgenic fishes will prove widely marketable and socially sustainable.

Empirical evaluations of both benefits and risks are yet to be conducted for other lines of transgenic fishes that may be considered for commercial production. Trade-offs among benefits and hazards may be anticipated. For example, as noted above, production of fish expressing a phytase transgene would take advantage of their ability to more efficiently utilize seed meals, increasing feed conversion efficiency and reducing the discharge of unutilized phosphorus from feed, decreasing eutrophication of waters in culture facilities and in waters receiving aquaculture discharges [41]. However, ability to efficiently utilize a wider range of natural foods might render escaped transgenic fish more competitive in accessible ecosystems. In particular, phytase-transgenic tilapia [42] might become even more competitive than wild-type tilapia, which are already recognized as invasive in a range of ecosystems in tropical and subtropical regions [98]. While effective confinement might be achievable in well-managed commercial operations, it likely cannot be achieved on subsistence farms, greatly influencing the assessment of whether phytase-transgenic tilapia would promote sustainability in developing countries.

As noted above, several transgenic fish lines have been produced for use as biomonitors [16–18] for detection of waterborne mutagens that would trigger expression of readily detected marker compounds in the fish. Such fishes would be exposed to such compounds by holding them in confinements in the waterbody of interest or by transporting the water to a confined fish-holding system. The benefits of long-term, integrative monitoring of water quality are clear. Any risks that might flow from the possibility of a nonindigenous species being used, escaping, and becoming naturalized in the waterbody at issue could be managed by use of effective confinement or by use of a species that cannot become established in the accessible ecosystem, for example, a tropical species sensitive to winter temperatures in the temperate zone.

These examples indicate how trade-offs of economic benefits and ecological risks, adoption of risk management measures, and ecological and social

context interact to affect the sustainability of transgenic fishes for aquaculture and other applications. Explicit consideration of benefit and ecological risk, and hopefully also ecological, economic, and social aspects of sustainability would accompany societal consideration of whether and how to go forward with commercial production of transgenic fishes.

Oversight of production of transgenic fishes. The decision of whether and under what conditions production of transgenic fish would go forward will largely be made at the national level. As noted above, at least three countries – the USA, Cuba, and the People’s Republic of China – are considering applications for commercial production of transgenic fish. The USA, Canada, the European Union, and Norway have regulatory systems in place for oversight of aquaculture biotechnology. Under Article 21 of the United Nations Convention on Environment and Development and the subsequent Cartagena Protocol, signatories commit to developing and implementing policies for oversight of biotechnology. Consequently, many countries – for example, Cuba, Thailand, and the People’s Republic of China [99] – are in the process of developing and implementing policy and staffing government offices that would be utilized to consider applications for production of transgenic fish.

Adaptive and proactive management. Many critical unknowns complicate risk assessment and risk management for transgenic aquaculture stocks. The adaptive management approach is based on recognition that knowledge of the environmental and social systems into which the aquaculture stocks would enter is *always* incomplete. Hence, management should evolve as knowledge of these systems increases [96]. Management cannot adapt if realized by a single passage through transgenesis and breeding, decision of whether and how to produce the transgenic stocks, and implementation of the distribution and production process. Instead, adaptive management would include risk assessment for candidate areas for distribution, incorporation of risk management into the production program, and capacity building as appropriate to meet program goals. Once transgenic stocks are distributed, culture operations and receiving ecosystems would be monitored for indicators of ecological and social conditions [96, 100]. Should monitoring indicate

that economic and social benefits are being realized without ecological harms occurring, then few if any adjustments to program implementation are required. However, should monitoring indicate that production of cultured stocks is *not* contributing to nutritional and economic well-being or that the stocks are escaping and impacting accessible ecosystems, then it will prove necessary to redefine goals, revise implementation, and continue monitoring.

The issue of whether and under what conditions transgenic fish would promote sustainability of aquaculture has been raised largely after research and development on the first wave of transgenic fish lines was well advanced, after proponents have applied for regulatory approval for commercial production. Kapuscinski et al. [101] proposed a proactive “safety-first” approach in which an early, prospective risk assessment would be conducted in order to guide planning and implementation of measures to prevent or minimize risk as development of a transgenic line progresses. Transgene constructs would be designed for safety. Gene transfer scientists would strive for better control of copy number, genomic insertion site, and control of expression of the transgene, contributing to better-controlled modification of phenotype. Transgenic lines would be evaluated for stability of transgenic expression and transmission. Development of inducible expression of transgenes would reduce risk should transgenic fish escape into the environment. Improvements in confinement also would reduce risk posed by aquaculture production of transgenic fishes.

Applications of transgenic fishes, the science of risk assessment, the practice of risk management, and public policies for oversight of biotechnology are all in development. The degree to which production of transgenic fishes ultimately will prove sustainable will depend upon many societal decisions of whether and under what conditions to utilize transgenic technology in aquaculture.

Bibliography

Primary Literature

1. Zbikowska HM (2003) Fish can be first – advances in fish transgenesis for commercial applications. *Transgenic Res* 12:379–389
2. Zhu Z, Li G, He L, Chen S (1985) Novel gene transfer into the fertilized eggs of goldfish (*Carassius auratus* L. 1758). *Z Angew Ichthyol* 1:31–34
3. Chourrout D, Guyomard R, Houdebine L-M (1986) High efficiency gene transfer in rainbow trout (*Salmo gairdneri* Rich.) by micro-injection into egg cytoplasm. *Aquaculture* 51:143–150
4. Inoue K, Yamashita S, Hata J, Kabeno K, Asada S, Nagahisa E, Fujita T (1990) Electroporation as a new technique for producing transgenic fish. *Cell Diff Dev* 29:123–128
5. Powers DA, Hereford I, Cole T, Chen TT, Lin CM, Kight K, Creech K, Dunham R (1992) Electroporation: a method for transferring genes into the gamete of zebrafish (*Brachiodanio rerio*), channel catfish (*Ictalurus punctatus*), and common carp (*Cyprinus carpio*). *Mol Mar Biol Biotechnol* 1:301–308
6. Muller F, Ivics Z, Erdelyi F, Papp T, Varadi I, Horvath L, Maclean N (1992) Introducing foreign genes into fish eggs with electroporated sperm as a carrier. *Mol Mar Biol Biotechnol* 1:276–281
7. Yamauchi M, Kinoshita M, Sasanuma M, Tsuji S, Terada M, Morimyo M, Ishikawa Y (2000) Introduction of a foreign gene into medakafish using the particle gun method. *J Exp Zool* 287:285–293
8. Ivics Z, Izsvak Z, Hackett PB (1993) Enhanced incorporation of transgenic DNA into zebrafish chromosomes by a retroviral integration protein. *Mol Mar Biol Biotechnol* 2:166–173
9. Izsvak Z, Ivics Z, Plasterk RH (2000) Sleeping beauty, a wide host-range transposon vector for genetic transformation in vertebrates. *J Mol Biol* 302:93–102
10. Stuart GW, McMurray JV, Westerfield M (1988) Replication, integration, and stable germ-line transmission of foreign sequences injected into early zebrafish embryos. *Development* 103:403–412
11. Dodd A, Curtis PM, Williams LC, Love DR (2000) Zebrafish: bridging the gaps between development and disease. *Hum Mol Genet* 9:2443–2449
12. Dooley K, Zon LI (2000) Zebrafish: a model system for the study of human disease. *Curr Opin Genet Dev* 10:252–256
13. Ward AC, Lieschke GJ (2002) The zebrafish as a model system for human disease. *Front Biosci* 7:827–833
14. Gross ML, Schneider JF, Moav N, Alvarez C, Myster SH, Liu Z, Hew C, Hallerman EM, Hackett PB, Guise KS, Faras AJ, Kapuscinski AR (1992) Molecular analysis and growth evaluation of northern pike (*Esox lucius*) microinjected with growth hormone genes. *Aquaculture* 103:253–273
15. Hallerman EM (2004) GloFish, the first GM animal commercialized, profits amid controversy. ISB (Information Systems for Biotechnology) news report, June 2004. www.isb.vt.edu/news/2004/artspdf/june0405.pdf, accessed November 22, 2010
16. Winn RN, Norris MB, Brayer KJ, Torres C, Muller SL (2000) Detections of mutations in transgenic fish carrying a bacteriophage λ cII transgene target. *Proc Nat Acad Sci USA* 97:12655–12660
17. Amanuma K, Takeda H, Amanuma H, Aoki Y (2000) Transgenic zebrafish for detecting mutations caused by compounds in aquatic environments. *Nat Biotechnol* 18:62–65

18. Carvan MJ III, Dalton TP, Stuart GW, Nebert DW (2000) Transgenic zebrafish as sentinels for aquatic pollution. *Ann NY Acad Sci* 919:133–147
19. Food and Agriculture Organization of the United Nations (2009) State of world fisheries and aquaculture. Fisheries and Aquaculture Department, FAO, Rome. <http://www.fao.org>, accessed October 30, 2009
20. Myers RA, Worm B (2003) Rapid worldwide depletion of predatory fish communities. *Nature* 423:280–283
21. Goldburg R, Triplett T (1997) Murky waters: environmental effects of aquaculture in the United States. Environmental Defense Fund, New York, p 198. www.edf.org/documents/490_AQUA.pdf, accessed November 22, 2010
22. Black KD (2001) Environmental impacts of aquaculture. CRC, Boca Raton, p 214
23. Bardach JE (ed) (1997) Sustainable aquaculture. Wiley, New York, p 251
24. Svennevig N, Reinertsen H, New M (eds) (1999) Sustainable aquaculture: food for the future? Balkema Publishers, Rotterdam, p 348
25. Jana BB, Webster CD (eds) (2003) Sustainable aquaculture: global perspectives. Haworth, New York, p 365
26. Gomes EF, Rema P, Kaushik SJ (1995) Replacement of fish meal by plant proteins in the diet of rainbow trout (*Oncorhynchus mykiss*): digestibility and growth performance. *Aquaculture* 130:177–186
27. Gatlin DM III, Barrows FT, Brown P, Dabrowski K, Gaylord TG, Hardy RW, Herman E, Hu G, Krogdahl A, Nelson R, Overturf K, Rust M, Sealey W, Skonberg D, Souza EJ, Stone D, Wilson R, Wurtele E (2007) Expanding the utilization of sustainable plant produce in aquafeed: a review. *Aquacult Res* 38: 551–579
28. Pierce LR, Palti Y, Silverstein JT, Barrows FT, Hallerman EM, Parsons JE (2008) Family growth response to fishmeal and plant-based diets shows genotype x diet interaction in rainbow trout (*Oncorhynchus mykiss*). *Aquaculture* 278:37–42
29. Dupont-Nivet M, Medale F, Leonard J, LeGuillou S, Tiquet F, Quillet E, Geurden I (2009) Evidence of genotype-diet interactions in the response of rainbow trout (*Oncorhynchus mykiss*) clones to a diet with or without fishmeal at early growth. *Aquaculture* 295:15–21
30. Du SJ, Gong ZY, Fletcher GL, Shears MA, King MJ, Idler DR, Hew CL (1992) Growth enhancement in transgenic Atlantic salmon by the use of an all-fish chimeric growth hormone gene construct. *Biotechnology* 10:176–181
31. Fletcher GL, Shears MA, Yaskowiak ES, King MJ, Goddard SV (2004) Gene transfer: potential to enhance the genome of Atlantic salmon for aquaculture. *Aust J Exp Agric* 44: 1095–1100
32. Devlin RH, Yesaki TY, Biagi CA, Donaldson EM, Swanson P, Chan WK (1994) Extraordinary salmon growth. *Nature* 371:209–210
33. Rahman MA, Ronyai A, Engidaw BZ, Jauncey K, Hwang GL, Smith A, Roderick E, Penman D, Varadi L, Maclean N (2001) Growth and nutritional trials on transgenic Nile tilapia containing an exogenous fish growth hormone gene. *J Fish Biol* 59:62–78
34. Hinitz Y, Moav B (1999) Growth performance studies in transgenic *Cyprinus carpio*. *Aquaculture* 173:285–296
35. Venugopal T, Anathy V, Kirankumar S, Pandian TJ (2004) Growth enhancement and food conversion efficiency of transgenic fish, *Labeo rohita*. *J Exp Biol* 301A:477–490
36. Nam YK, Noh JK, Cho YS, Cho HJ, Cho KN, Kim CG, Kim DS (2001) Dramatically accelerated growth and extraordinary gigantism of transgenic mud loach *Misgurnus mizolepis*. *Transgenic Res* 10:353–362
37. Hew CL, Poon R, Xiong F, Gauthier S, Shears M, King M, Davies P, Fletcher G (1999) Liver-specific and seasonal expression of transgenic Atlantic salmon harboring the winter flounder antifreeze protein gene. *Transgenic Res* 8:405–414
38. Dunham RA, Warr GW, Nichols A, Duncan PL, Argue B, Middleton D, Kucuktas H (2002) Enhanced bacterial disease resistance of transgenic channel catfish *Ictalurus punctatus* possessing cecropin genes. *Mar Biotechnol* 4:338–344
39. Zhong J, Wang Y, Zhu Z (2002) Introduction of the human lactoferrin gene into grass carp (*Ctenopharyngodon idellus*) to increase resistance against GCH virus. *Aquaculture* 214: 93–101
40. Mao W, Wang Y, Wang W, Wu B, Feng J, Zhu Z (2004) Enhanced resistance to *Aeromonas hydrophila* infection and enhanced phagocytic activities in human lactoferrin-transgenic grass carp (*Ctenopharyngodon idellus*). *Aquaculture* 242:93–103
41. Hostetler HA, Collodi P, Devlin RH, Muir WM (2005) Improved phytate phosphorus utilization by Japanese medaka transgenic for the *Aspergillus niger* phytase gene. *Zebrafish* 2:19–31
42. Kemeh H (2004) Development of phytase transgenic Nile tilapia. Ph.D. dissertation, Purdue University, West Lafayette. www.docs.lib.purdue.edu, accessed October 30, 2009
43. Wong AC, Van Eenennaam AL (2008) Transgenic approaches for the reproductive containment of genetically engineered fish. *Aquaculture* 275:1–12
44. Uzbekova S, Chyb J, Ferriere F, Bailhache T, Prunet P, Alestrom P, Breton B (2000) Transgenic rainbow trout expressed sGnRH-antisense RNA under the control of sGnRH promoter of Atlantic salmon. *J Mol Endocrinol* 25:337–350
45. Thresher R, Grewe P, Patil J, Whyhard S, Templeton M, Chaimongkol A, Hardy C, Hinds L, Dunham R (2009) Development of repressible sterility to prevent the establishment of feral populations of exotic and genetically modified animals. *Aquaculture* 290:104–109
46. Krasnov A, Pitkanen TI, Molsa H (1999) Gene transfer for targeted modification of salmonid fish metabolism. *Genet Anal* 15:115–119
47. Guillen I, Berlanga J, Valenzuela CM, Morales A, Toledo J, Estrada MP, Puentes P, Hayes O, LaFuente J (1999) Safety evaluation of transgenic tilapia with accelerated growth. *Mar Biotechnol* 1:2–14
48. Wu G, Sun Y, Zhu Z (2003) Growth hormone gene transfer in common carp. *Aquat Living Resour* 16:416–420

49. National Research Council (2002) Animal biotechnology: science-based concerns. National Academy, Washington, DC, p 179. <http://www.nap.org>, accessed November 22, 2010
50. Pew Initiative on Food and Biotechnology (2003) Future fish: issues in science and regulation of transgenic fish. www.pewtrusts.org/our_work_report_detail.aspx?id.33350, accessed November 22, 2010
51. Food and Agriculture Organization of the United Nations and World Health Organization (2004) Safety assessment of foods derived from genetically modified animals, including fish. FAO food and nutrition paper 79, FAO, Rome
52. Hallerman EM, McLean E, Fleming IA (2006) Effects of growth hormone transgenes on the behavior and welfare of aquacultured fishes: a review identifying research needs. *Appl Anim Behav Sci* 104:265–294
53. Schiermeier Q (2003) Fish farms' threat to salmon stocks exposed. *Nature* 425:753
54. Volpe J, Taylor E, Rimmer D, Glickman B (2000) Evidence of natural reproduction of aquaculture-escaped Atlantic salmon in a coastal British Columbia River. *Conserv Biol* 14:899–903
55. Soto D, Jara F, Moreno C (2001) Escaped salmon in the inner seas, southern Chile: facing ecological and social conflicts. *Ecol Appl* 11:1750–1762
56. Council on Environmental Quality and Office of Science and Technology Policy (2000) Case study no. 1: growth-enhanced salmon. <http://www.ostp.gov/what'snew/casestudies>, accessed January 20, 2000
57. Schramm HL Jr, Piper RG (eds) (1995) Uses and effects of cultured fishes in aquatic ecosystems. In: American fisheries society symposium 15, American Fisheries Society, Bethesda
58. McGinnity P, Prodohl P, Ferguson A, Hynes R, Maoileidigh NO, Baker N, Cotter D, O'Hea B, Cooke D, Rogan G, Taggart J, Cross T (2003) Fitness reduction and potential extinction of wild populations of Atlantic salmon, *Salmo salar*, as a result of interactions with escaped farmed salmon. *Proc R Soc Lond B* 270:2443–2450
59. Fleming IA, Hindar K, Mjølnerød IB, Jonsson B, Balstad T, Lamberg A (2000) Lifetime success and interactions of farm salmon invading a native population. *Proc R Soc Lond B* 267:1517–1523
60. Nickum, MJ, Mazik PM, Nickum JG, MacKinlay DD (eds). (2004) Propagated fish in resource management. In: American fisheries society symposium 44, American Fisheries Society, Bethesda, MD
61. Verspoor E, Stradmeyer L, Nielsen JL (eds) (2007) The Atlantic salmon: genetics, conservation, and management. Blackwell, Oxford
62. Gross M (1998) One species with two biologies: Atlantic salmon (*Salmo salar*) in the wild and in aquaculture. *Can J Fish Aquat Sci* 55:131–144
63. Saegrov H, Hindar K, Kalas S, Lura H (1997) Escaped farmed Atlantic salmon replace the original salmon stock in the River Vosso, western Norway. *ICES J Mar Sci* 54:1166–1172
64. Kapuscinski AR, Hallerman EM (1990) Transgenic fish and public policy. Anticipating environmental impacts of transgenic fish. *Fisheries* 15(1):2–11
65. Kapuscinski AR, Hallerman EM (1991) Implications of introduction of transgenic fish into natural ecosystems. *Can J Fish Aquat Sci* 48(Suppl 1):99–107
66. Hallerman EM, Kapuscinski AR (1992) Ecological implications of using transgenic fishes in aquaculture. *ICES Mar Sci Symp* 194:56–66
67. Hallerman EM, Kapuscinski AR (1993) Potential impacts of transgenic and genetically manipulated fish on wild populations: addressing the uncertainties through field testing. In: Cloud JG, Thorgaard GH (eds) Genetic conservation of salmonid fishes. Plenum, New York, pp 93–112
68. Tiedje JM, Colwell RK, Grossman YL, Hodson RE, Lenski RE, Mack RN, Regal PJ (1989) The planned introduction of genetically engineered organisms: ecological considerations and recommendations. *Ecology* 70:298–315
69. Snow A, Andow D, Gepts P, Hallerman E, Powers A, Tiedje J, Wolfenbarger L (2005) Ecological risks and benefits associated with the environmental release of GMOs. *Ecol Appl* 15:377–404
70. Stevens ED, Sutterlin A, Cook T (1998) Respiratory metabolism and swimming performance in growth hormone transgenic Atlantic salmon. *Can J Fish Aquat Sci* 55:2028–2035
71. Cook JT, McNiven MA, Sutterlin AM (2000) Metabolic rate of pre-smolt growth-enhanced transgenic Atlantic salmon (*Salmo salar*). *Aquaculture* 188:33–45
72. Abrahams MV, Sutterlin A (1999) The foraging and antipredator behavior of growth-enhanced transgenic Atlantic salmon. *Anim Behav* 58:933–942
73. Devlin RH, Johnsson JI, Smailus DE, Biagi CA, Johnsson E, Björnsson BT (1999) Increased ability to compete for food by growth hormone transgenic coho salmon (*Oncorhynchus kisutch* Walbaum). *Aquacult Res* 30:1–4
74. Devlin RH, Biagi CA, Yesaki TY (2004) Growth viability and genetic characteristics of GH transgenic coho salmon strains. *Aquaculture* 236:607–632
75. Sundstrom LF, Lohmus M, Johnsson JI, Devlin RH (2004) Growth hormone transgenic salmon pay for growth potential with increased predation mortality. *Proc R Soc Lond B* 271:S350–S352
76. Duan M, Zhang T, Hu W, Sundstrom LF, Wang Y, Li Z, Zhu Z (2009) Elevated ability to compete for limited food resources by "all-fish" growth hormone transgenic common carp (*Cyprinus carpio* L.). *J Fish Biol* 75:1459–1469
77. Devlin RH, Sundstrom LF, Johnsson JI, Fleming IA, Hayes KR, Ojwang WO, Bamaradeniya C, Zakaria-Ismail M (2007) Assessing ecological effects of transgenic fish prior to entry into nature. In: Kapuscinski AR, Hayes KR, Li S, Dana G, Hallerman EM, Schei PJ (eds) Environmental risk assessment of genetically modified organisms, vol 3, Methodologies for transgenic fish. CAB International, Wallingford, Oxfordshire, pp 151–187
78. National Research Council (2004) Biological confinement of genetically engineered organisms. National Academy, Washington, DC, p 255. <http://www.nap.org>, accessed November 22, 2010
79. Devlin RH, Biagi CA, Yesaki TY, Smailus DE, Byatt JC (2001) A growth-hormone transgene boosts the size of wild but not domesticated trout. *Nature* 409:781–782

80. Devlin RH, D'Andrade M, Uh M, Biagi CA (2004) Population effects of growth hormone transgenic coho salmon depend on food availability and genotype by environment interactions. *Proc Nat Acad Sci USA* 101: 9303–9308
81. Bessey C, Devlin RH, Liley NR, Biagi CA (2004) Reproductive performance of growth-enhanced transgenic coho salmon. *Trans Am Fish Soc* 133:1205–1220
82. Muir WM, Howard RD (1999) Possible ecological risks of transgenic organism release when transgenes affect mating success: sexual selection and the Trojan gene hypothesis. *Proc Nat Acad Sci USA* 96:13853–13856
83. Muir WM, Howard RD (2001) Fitness components and ecological risks of transgenic release: a model using Japanese medaka (*Oryzias latipes*). *Am Nat* 158:1–16
84. Muir WM, Howard RD (2002) Assessment of the possible ecological risks and hazards of transgenic fish with implications for other sexually reproducing organisms. *Transgenic Res* 11:101–114
85. Hedrick PW (2001) Invasion of transgenes from salmon or other genetically modified organisms into natural populations. *Can J Fish Aquat Sci* 58:841–844
86. Kapuscinski AR, Hard JJ, Paulson KM, Neira R, Ponniah A, Kamonrat W, Mwanja W, Fleming IA, Gallardo J, Devlin RH, Tresiak J (2007) Approaches to assessing gene flow. In: Kapuscinski AR, Hayes KR, Li S, Dana G, Hallerman EM, Schei PJ (eds) *Environmental risk assessment of genetically modified organisms*, vol 3, Methodologies for transgenic fish. CAB International, Wallingford, Oxfordshire, pp 112–150
87. Agricultural Biotechnology Research Advisory Committee – US Department of Agriculture (1995) Performance standards for safely conducting research with genetically modified fish and shellfish. <http://www.isb.vt.edu/perfstands>, accessed October 30, 2009
88. Mair GC, Nam YK, Solar II (2007) Risk management: reducing risk through confinement of transgenic fish. In: Kapuscinski AR, Hayes KR, Li S, Dana G, Hallerman EM, Schei PJ (eds) *Environmental risk assessment of genetically modified organisms*, vol 3, Methodologies for transgenic fish. CAB International, Wallingford, Oxfordshire, pp 209–238
89. Bye VJ, Lincoln RF (1986) Commercial methods for the control of sexual maturation of rainbow trout (*Salmo gairdneri*). *Aquaculture* 57:299–309
90. Mair GC, Abucay JS, Skibinski DOF, Abella TA, Beardmore JA (1997) Genetic manipulation of the sex ratio for large scale production of all-male tilapia *Oreochromis niloticus* L. *Can J Fish Aquat Sci* 54:396–404
91. Dunham RA (2004) *Aquaculture and fisheries biotechnology: genetic approaches*. CABI, UK, p 372
92. Dolin EJ (2003) *Snake head: a fish out of water*. Smithsonian Books, Washington, DC
93. Kapuscinski AR, Li S, Hayes K, Dana G, Hallerman E, Schei P (eds) (2007) *Environmental risk assessment of genetically modified organisms*, vol 3, Methodologies for transgenic fish. CABI Publishers, Wallingford, Oxfordshire
94. Hayes KR, Kapuscinski AR, Dana G, Li S, Devlin RH (2007) Introduction to environmental risk assessment for transgenic fish. In: Kapuscinski AR, Hayes KR, Li S, Dana G, Hallerman EM, Schei PJ (eds) *Environmental risk assessment of genetically modified organisms*, vol 3, Methodologies for transgenic fish. CAB International, Wallingford, Oxfordshire, pp 1–28
95. National Research Council (1996) *Understanding risk: informing decisions in a democratic society*. National Academy, Washington, DC
96. Kapuscinski AR, Nega T, Hallerman EM (1999) Adaptive biosafety assessment and management regimes for aquatic genetically modified organisms in the environment. In: Pullin RSV, Bartley DM, Kooiman J (eds) *Towards policies for conservation and sustainable use of aquatic genetic resources*. In: ICLARM (International Center for Living Aquatic Resources Management) conference proceedings 59, pp 225–251
97. Rollin BE (1995) *The Frankenstein syndrome: ethical and social issues in the genetic engineering of animals*. Cambridge University Press, Cambridge, UK
98. Moreau J (1983) A review of introductions of tilapia in open waters of Africa, their influence on ecology and fisheries. In: Fishelson L, Yaron Z (eds) *International symposium on tilapia in aquaculture*, Tel Aviv University Press, Nazareth, Israel, May 8–13 1983, pp 77–85
99. Nelson KC, Basiao Z, Cooper AM, Dey M, Foniciella D, Lorenzo Hernandez M, Kunawasen S, Leelapatra W, Li S, Ratner BD, Toledo MI (2007) Problem formulation and options assessment: science-guided deliberation in environmental risk assessment of transgenic fish. In: Kapuscinski AR, Hayes KR, Li S, Dana G, Hallerman EM, Schei PJ (eds) *Environmental risk assessment of genetically modified organisms*, vol 3, Methodologies for transgenic fish. CAB International, Wallingford, Oxfordshire, pp 29–60
100. Senenan W, Hard JJ, Alcivar-Warren A, Trisak J, Zakaria-Ismael M, Lorenzo Hernandez M (2007) Risk management: post-approval monitoring and remediation. In: Kapuscinski AR, Hayes KR, Li S, Dana G, Hallerman EM, Schei PJ (eds) *Environmental risk assessment of genetically modified organisms*, vol 3, Methodologies for transgenic fish. CAB International, Wallingford, Oxfordshire, pp 239–271
101. Kapuscinski AR, Goodman RM, Hann SD, Jacobs LR, Pullins EE, Johnson CS, Kinsey JD, Krall RL, La Vina A, Mellon M, Ruttan V (2003) Making “safety first” a reality for biotechnology products. *Nature Biotech* 21:599–601

Books and Reviews

- Beaumont AR, Hoare K (2003) *Biotechnology and genetics in fisheries and aquaculture*. Blackwell, Oxford, p 158
- Bondad-Reantaso MG, Arthur JR, Subasinghe RP (eds) (2008) *Understanding and applying risk analysis in aquaculture*. FAO fisheries technical paper, no 519. FAO, Rome

- Devlin RH, Sundstrom LF, Muir WM (2006) Interface of biotechnology and ecology for environmental risk assessments of transgenic fishes. *Trends Biotechnol* 24:89–97
- Ennion SJ, Goldspink G (eds) (1996) Gene expression and manipulation in aquatic organisms. Cambridge University Press, Cambridge, UK, p 214
- Hew CL, Fletcher GL (eds) (1992) Transgenic fish. World Scientific Publishing Co., Singapore
- Hew CL, Fletcher GL (2001) The role of aquatic biotechnology in aquaculture. *Aquaculture* 197:191–204
- Iyengar A, Muller F, Maclean N (1996) Regulation and expression of transgenes in fish – a review. *Transgenic Res* 5:147–166
- Maclean N, Laight RJ (2000) The application of gene manipulation to aquaculture. *Aquaculture* 85:1–20
- McLean E, Devlin RH (2000) Application of biotechnology to enhance growth of salmonids and other fish. In: Fingerma M, Nagabhushanam R (eds) Recent advances in marine biotechnology, vol 4, Aquaculture, Part B: Fishes. Science Publishers, Enfield, pp 17–55
- Mench JA (1999) Ethics, animal welfare and transgenic farm animals. In: Murray JD, Anderson GB, Oberbauer AM, McGlothin MM (eds) Transgenic animals in agriculture. CAB International, Wallingford, pp 251–268
- Millar K, Tomkins S (2007) Ethical analysis of the use of GM fish: emerging issues for aquaculture. *J Agr Environ Ethic* 20:437–453
- Organization for Economic Cooperation and Development (1995) Environmental impacts of aquatic biotechnology. OECD, Paris
- Rasmussen RS, Morrissey MT (2007) Biotechnology in aquaculture: transgenics and polyploidy. *Comp Rev Food Sci Food Safety* 6:2–16

Transgenic Livestock for Food Production, Introduction

C. BRUCE A. WHITELAW

The Roslin Institute and Royal (Dick) School of Veterinary Studies, Division of Developmental Biology, University of Edinburgh, Scotland, UK

Our societies currently face many big challenges: how to feed the increasing world population, how to balance the competing demands for land and water, how to mitigate climate change, and how to maintain healthy lifestyle into old age. Among other technologies, the animal genetic revolution offers several strategies to mitigate these challenges and provide benefit for mankind. While achieving this, we need to maintain and where appropriate increase the welfare and existence of animals we use in agriculture and biomedicine. Transgenic

technologies, the ability to add or transfer or change the genetic makeup of an animal, can contribute. In the following ten entries, this premise is presented in detail. Why the entries were selected and how they complement each other while giving an overview of the field of transgenic animal biotechnology is described below.

The cornerstone of biological research is the ability to experimentally alter gene activity. In this way, the functional importance of a gene can be studied. This can be achieved indirectly through altering the environment of cells or organisms, for example, adding specific nutrients to a culture condition or infecting with a pathogen. Alternatively, direct manipulation can be achieved through genetic engineering. This section focuses on the later strategy which includes gene addition, gene removal (or inactivation), or gene alteration. To illustrate this, gene addition might be appropriate where an enhanced functionality is required such as food utilization, while gene inactivation would be appropriate if the aim was to remove a food allergen. Both have been successfully applied to animals and this is likely to be so going forward. In addition, as methodology becomes more sophisticated and the ability to directly alter a given gene sequence, this strategy will see more and more application; for example, where gene diversity is identified that could be applied to engineer altered metabolism or reduce disease susceptibility.

Although initially performed in bacteria, researchers quickly developed the tools to genetically modify animals. First in mice at the end of the 1970s, then livestock in the following decade, in parallel strategies for altering the germ line in fish and birds were developed. Researchers have developed strategies which enable precise spatial and temporal changes to target gene activity. Many consider transgenic or genetic modification (GM) methods (► [Transgenics: Alternative Gene Transfer Methods](#)) as part of the technical continuum that is available to animal breeders. As such, GM or transgenic animal technology would simply represent new methodology alongside the more established tools as artificial insemination and embryo transfer. These new animal reproductive tools include somatic cloning and in vitro embryo production, and when combined with the emerging molecular genetic tools provide powerful new opportunities for livestock breeding.

The first transgenic livestock were reported in 1985. Since these early days of this technology, there has been

a steady increase in the availability of tools available (► [Transgenic Livestock Technologies](#)) for this form of assisted reproduction in animals. Some developments have been heralded as breakthroughs only to quickly drop out of favor while others have received limited use. Recently, reagents based on precisely targeted enzymatic cleavage of the genome that in addition to strategies such as gene knockdown provide the animal geneticist with both efficient and sophisticated methods with which to generate transgenic animals.

Perhaps one of the most consequential scientific advances of recent years was the technical breakthrough that produced “Dolly” the sheep at the Roslin Institute. This animal was produced by nuclear transfer (also termed SCNT, somatic cell nuclear transfer), or as it is commonly referred to as cloning. Animal cloning involves the replacement of the genetic material found in an unfertilized egg with that from a donor cell’s genetic material. SCNT reflects considerable scientific thinking and offers the experimental biologist considerable insights into the functional reprogramming of cellular function in addition to details of cellular subcomponent roles in normal cell activity (► [Livestock Somatic Cell Nuclear Transfer](#)). In addition to being a valuable aspect of research in its own right, SCNT can be combined with genetic engineering tools to precisely modify the genome of mammals (► [Nuclear Transfer to Produce Transgenic Mammals](#)). This combination offers exciting commercial opportunities across the fields of agriculture and animal breeding, to the biomedical arena and novel biotechnological processes.

Transgenic technologies are not restricted to mammals. Due to its use in food production and as an experimental organism in scientific research, the chicken is the most widely used bird species for experimental development of avian transgenic technologies (► [Avian Specific Transgenesis](#)). Perhaps the most exciting recent advance is that of precise gene modification in special avian cells called primordial stem cells. These new tools offer strategies for the production of transgenic birds other than the chicken. As such, this is an area currently experiencing rapid development and many expect it will be in birds that GM technologies will be first successfully exploited to produce animals that have enhanced resistance to infectious pathogens.

Although it can be argued that technically transgenic technology is most advanced for mammals, and

currently developing quickest for birds, it is considered very likely that the first commercial application of transgenic animals that will enter the food chain will be in fish (► [Transgenic Fishes: Applications, State of the Art, and Risk Concerns](#)). Fish that grow faster are providing the first test of the regulatory systems while providing the focus for much debate on broader and longer-term questions regarding whether such use would prove ecologically, economically, and socially sustainable. There is also considerable effort being directed at the development of transgenic strategies to combat disease in fish populations, and the technology can be equally applied to food fish as to sport fish.

As the world population increases there is an increasing demand for food including both plant and animal products. The demand for meat and milk are increasing at a faster rate than for plant products because of increased wealth in developing countries. Animal products represent a concentrated protein and vitamin source which complements cereal and other vegetable proteins. Intensification of livestock production to satisfy the demand is exacerbating the deleterious impact of intensive animal agriculture on the environment and new approaches are needed to reduce the impact. Transgenic animals that have a smaller environmental footprint, increased productivity brought about by enhancing aspects of agriculturally important traits can contribute to this societal challenge. For this potential to be realized an active dialogue between all stakeholders is required to achieve ethically accepted sustainable future animal production.

The animals we all recognize in farms today are the result of many centuries of domestication. The driving force for this has largely come from man’s desire for meat. This single process has shaped the rural landscape we all recognize today. It is therefore not surprising that animal breeders look to transgenic strategies to enhance the traditional animal breeding traits including food production (► [Transgenic Livestock, Enhanced Nutritional Quality in](#)). The focus now is not on quantity but on quality, while reflecting the rapid industrialization of our world and the associated increase in affluence experienced by many (but not all) of our communities. Food choice is now considered alongside sustainability. Animal welfare is paramount while still recognizing that many of the world

population live in poverty with starvation a constant threat. Transgenic technologies offer the potential of quick solutions to the production of food products with enhanced specific nutritional characteristics from animals. The goal is healthier and safer food. However, this potential will only be realized with consumer acceptance in GM technology.

Two specific examples of traditional breeding traits that transgenic technology may provide innovative solutions are worthy of specific attention. The first is that of animal fertility (► [Transgenic Technologies and Increased Livestock Fertility](#)). Reproductive efficiency of domestic animals is critical to the sustainability of modern livestock industries. With various indicators of reducing fertility now appearing in our farms it is timely to consider innovative approaches to halt this decline. Aspects under investigation include increasing litter size in livestock or egg-laying capacity in poultry. More radically this technology could be applied to reduce the limitations associated with seasonal breeding.

The second topic that is attracting considerable interest in both the scientific research community and in the animal breeding industry is that of animals that are less susceptible to the ravages and distress of disease (► [Disease-Resistant Transgenic Animals](#)). Offering both enhanced animal welfare and economic advantages this field is both fast moving and innovative, building much on the genomic revolution which is in turn constructed from the strong discipline of genetics. With huge impact on global animal health and the associated impact of zoonotic disease in man, this is an area which offers huge benefit to man and animal alike.

The simultaneous need for more food as the world population expands and that of limiting both short-term and longer-term environmental consequences is a huge challenge for man. Increased animal productivity could directly compromising our environment. Transgenic technology through providing animals that have a smaller environmental footprint (► [Transgenic Livestock, Decreasing Environmental Impact of](#)) could form an important part of the compromise that is inevitable if we are to provide sufficient food for our increasing human needs. Under development are projects aiming to reduce manure output and decrease greenhouse gas production. It is argued that the food security agenda being promoted by many governments will lead to greater acceptance of transgenic animals,

following the slow but every increasing presence and consumption of transgenic plant products.

Transgenic animals present a dilemma for us. This technology offers rapid change with the promise of economic benefit, enhanced food security and better animal welfare. Yet transgenic technologies directly challenge many ethical aspects of thinking (► [Transgenic Livestock, Ethical Concerns and Debate](#)). Central to the process of acceptance is a conscious dialogue between all stakeholders – to build up a generally acceptable stance on what is good with respect to the ethical limits of human use of animals.

At the time of writing, there are no transgenic livestock in production – with faster growing fish being the closest and disease resistant birds being the most exciting. By the time of reading, this may well have changed – or be close to such a scenario. Agricultural advances have been enormously successful in providing an inexpensive supply of high-quality and safe foods. The new advance of transgenesis is likely to continue this trend providing some of the solutions to tomorrow's agricultural challenges. Governmental and industrial financial support – ideally in equal measures – is needed to achieve this. Equally important is the public dialogue to expose the benefits this technology can offer all stakeholders in our diverse world society.

Transgenic Livestock Technologies

CESARE GALLI^{1,2}, ANDREA PEROTA¹, GIOVANNA LAZZARI¹, FRANCO LUCCHINI³

¹Avantea, Laboratorio di Tecnologie della Riproduzione, Cremona, Italy

²Clinical Veterinary Department, University of Bologna, Italy

³Centro Ricerche Biotecnologiche, Università Cattolica del Sacro Cuore, Cremona, Italy

Article Outline

Glossary

Definition of the Subject

Introduction

Assisted Reproduction Techniques

Techniques for Genetic Engineering
 Transferring the DNA into the Cell
 Emerging Technologies
 Future Directions
 Acknowledgments
 Bibliography

Glossary

Genetic engineering Recombinant DNA technology, gene manipulation including experimental manipulation of genetic material for industrial or medical use.

Germ cells Cells able to give rise to a functional gamete in multicellular animals.

Assisted reproduction The medical and laboratory processes used to reproduce the steps necessary to generate a newborn starting from gametes and/or embryos.

IVF In Vitro Fertilization, the laboratory process in which a mammalian egg is fertilized outside the body and then put back inside to grow into a new individual.

ICSI Intra Cytoplasmic Sperm Injection, a procedure by which a spermatozoa is injected directly into the cytoplasm thus overcoming artificially the natural barriers of the oocyte.

SCNT Somatic Cell Nuclear Transfer, a technique that generates genomic copies of any given individual animal by replacing inside an oocyte its genome with the genome taken from somatic cells of a donor animal.

iPS cells Induced Pluripotent Stem cells, somatic cells that have been turned into pluripotent stem cells by the overexpression of pluripotency genes like OCT4, Nanog, Sox3, and Klf4.

Preimplantation embryo Following fertilization, the mammalian one-cell embryo (*zygote*) is still enclosed into the transparent proteic shell of the oocytes (*Zona pellucida*). During the journey along the oviduct it repeatedly cleaves to form an aggregate of smaller cells (*Morula Stage*) and finally develops an inner cavity (*Blastocyst stage*). Once in the uterus, the expanded blastocyst hatches breaking the zona pellucida and begins the implantation on the endometrium.

Chimerism The status of an organism made up by cells of different genotypes (like in the mythological monster *Chimera*). Chimeras derive from the aggregation of two different embryos of the same species or from two different species closely related (e.g., sheep and goat).

Mosaicism The cells of the organism own the same background genotype but are diverse in carrying or not a single mutation (or transgene). Such different cells may show a “patchy” distribution within tissues (like in a mosaic).

cdNA (complementary DNA) is a sequence of DNA obtained by reverse transcription of the mRNA (messenger RNA). It retains the whole information for coding the protein, but lacks the introns originally present in the chromosomal gene.

Zinc finger nuclease Recombinant enzyme able to bind a specific sequence of DNA and to “cut” it, generating a Double Strand Break (DSB). It is composed by an artificial DNA binding domain (a polypeptide with a repeat of regions each folded around a zinc ion, to evoke the idea of the fingers) linked to a bacterial nuclease domain, capable to hydrolyze the phosphate bond of the DNA.

Transposon A mobile genetic element able to transpose, that is, to excise itself from one site of the genome and to integrate in a different site. It does not move from one cell to another.

Retroviruses A class of viruses with a RNA genome, able to bind specific receptor on the target cell, to enter across the cell membrane and to transcribe its RNA in a DNA copy by means of its own RNA-dependent DNA polymerase (reverse transcriptase). Specific sequences promote the integration of the viral DNA into the genome of the host.

AAV A defective virus, with a small, single-stranded DNA genome, belonging to the family of Parvoviridae. AAV is able to replicate only in case of concomitant presence of an active adenovirus in the same cell. AAV can often integrate a double-stranded form of its DNA into the genome of the host.

Recombinase A family of enzymes able to exchange stretches of DNA among different molecules. Recombinases like CRE, coded by the genome of bacteriophage P1 (a virus parasiting a bacterium),

recognize specific sequences as target of their activity. In the presence of such Recombination Recognition Sites, phagic recombinase is able to insert and excise the viral genome from the bacterial chromosome.

siRNA Silencer RNA or short interfering RNA are small synthetic RNA molecules able to trigger the degradation of a target messenger RNA bearing the same sequence. This results in a posttranscriptional inhibition of gene expression. SiRNA complexes derive from the processing of artificial shRNAs (short hairpin RNA) bearing a secondary double-stranded structure that is subsequently processed by the endogenous cell machinery. RNA-mediated inhibition of gene expression, also known as RNA-interference, is naturally observed in most eukaryotes and relies on a specific pathway involving specific endoribonucleases and endogenous micro-RNA (miRNAs) genes.

Definition of the Subject

Recombinant DNA technology and gene transfer were set about 40 years ago and since then have provided powerful tools for the production of biological molecules, the development of new varieties of crops, and for generating key animal models in biomedical research.

Following the synthesis of recombinant insulin, many other pharmacological proteins have been produced by expressing heterologous genes in microorganisms, plants, and animals.

The application of gene transfer or “transgenesis” to animals has been developed mainly in mouse, while its use for generating transgenic livestock has encountered unanticipated resistances originated by social, economical and technical consideration. Apart from public concern and social consideration about the release of Genetically Engineered (GE) products in the market and GE animals into the environment, which will not be discussed in this chapter, the development of transgenic livestock has been slow due to the technical difficulties faced in transferring this technology from mouse to large domestic animals and by the initial discouraging results, outlining the difficulties in obtaining the expected effect upon a genetic modification.

This review will survey the scientific advancement making the generation of GE swine and ruminants

more accessible in the last few years, positively influencing the technical and, as a consequence, the economical issues related to their development.

Introduction

Genetic mutations and equally engineering (GE) of living organisms are part of biological processes that are inherited through reproduction across generations. Such events provide the genetic variability that is at the core of evolution of the species and ensure its survival both in plants and animal kingdoms. The understanding of the biology at the basis of such physiological events, the discovery of the mechanisms and the molecules involved has allowed the investigators to exploit and adapt such tools for the artificial modification of the genome of living organisms from virus and bacteria to mammals including primates. These developments that led to definitions and guidelines of what is now called DNA recombinant technology can be traced back to 1975 to the Asilomar Conference where the scientific community met to discuss potential hazard and regulation. DNA recombinant technology had the potential to be taken from bacteria and viruses to plants and animals with, at the time, unknown biological consequences and biosafety concerns. It was not too long in fact that the first GE mammals were produced [1, 2]. Palmiter's results were dramatic in the sense that the mouse engineered to express the rat growth hormone grew to the size double that of the normal mice, thus opening the door to the application of such technologies in livestock, as stated in the abstract “This approach has implications for studying the biological effects of growth hormone, as a way to accelerate animal growth, as a model for gigantism, as a means of correcting genetic disease, and as a method of farming valuable gene products”.

Since then, GE technology in mammals has been developed mainly in the laboratory mouse to study development and disease processes for research and clinical purposes with the ability to manipulate any given gene of the mouse genome. The developments in livestock immediately followed [3]; however it soon became evident that it was only the beginning and to reach such objective it was a difficult task and more knowledge was required. Beside the know-how required to engineer the genome, to make those modifications into the animals inherited through generations, it is necessary

to operate during the early stages of embryonic development or on the germ cells to be able to have all the cells originating from the embryo GE. This is the reason why the developments of assisted reproduction techniques have been crucial for the generation of transgenic animals, especially in livestock where the access *in vivo* to preimplantation embryos is more complex, expensive, and cumbersome than in the mouse. It can be stated that the developments in transgenic livestock have been set by the development of assisted reproduction techniques like oocyte *in vitro* maturation and fertilization, intracytoplasmic sperm injection, embryo culture, and somatic cell nuclear transfer (SCNT).

The injection of recombinant DNA into the pronucleus of the zygote has been the first technique used to generate GE livestock. This required a large supply of zygotes that can be easily generated in mice *in vivo* but are very expensive or sometimes impossible to obtain from the donor females of larger animals because of the costs involved and the low efficiency (only 1–5% of the injected embryos result in transgenic offspring).

Nevertheless, first studies on sheep and pigs involved the use of *in vivo* produced zygotes, adapting the protocols already set for the mouse [1, 3–5]. On the other hand, the generation of transgenic calves was not practically manageable by this route and could only be achieved [6] using *in vitro* maturation and fertilization of oocytes collected at the slaughterhouse [7] where large numbers can be generated at a low cost without the use of live animals.

Besides the classical microinjection procedure, alternative approaches have been developed to bring exogenous DNA into the nucleus of the embryo. Embryonic stem cells are largely used in the mouse to generate transgenic animals. They can be replicated in culture for a long period without losing their pluripotency, that is, their ability to take part in the formation of a developing embryo. Thanks to this peculiar behavior, they can be GE to obtain cell clones bearing specific modification of the genome. Such cells can be introduced in the early embryo where they participate in the development of a chimeric organism and eventually integrate into the germ line [8]. This approach has opened the way to the development of the so-called gene targeting, allowing the generation of transgenic mice bearing specific mutations, including

disruption of specific genes (knock-out) or insertion of new coding sequences in specific loci (knock-in). ES cells are not yet available in livestock, therefore this route is not yet practicable [9].

An alternative approach has been disclosed by the development of somatic cell nuclear transfer (SCNT) technology, currently the preferred route to make transgenic livestock [10]. Somatic cells may be GE, even by gene targeting, and then used for SCNT. Although SCNT is not very efficient, all the animals obtained carry the selected mutation that has been introduced in the somatic cells subsequently used for nuclear transfer.

In addition, alternative approaches have been developed to generate transgenic animals like the infection with an engineered retrovirus of *in vitro* cultured preimplantation embryos or the injection of a sperm together with the desired recombinant DNA vector during intracytoplasmic sperm injection (ICSI).

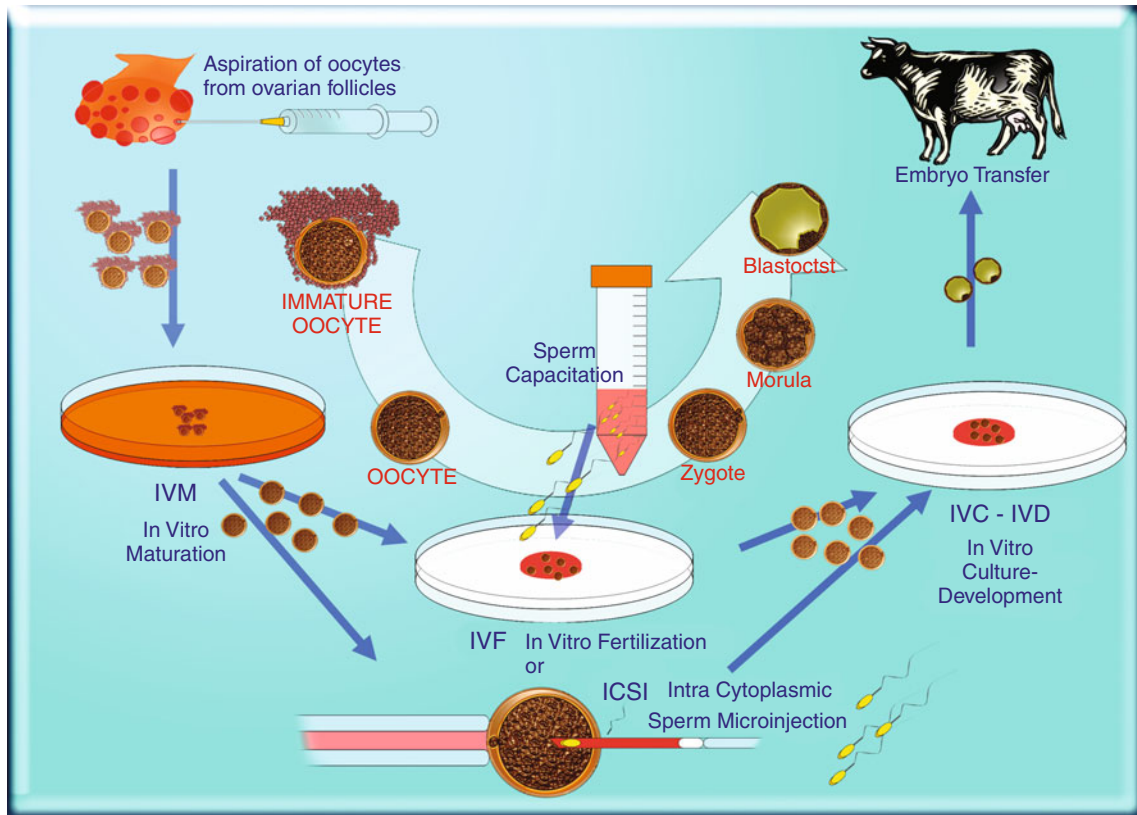
In the following paragraphs, the developments in assisted reproduction technologies that have made GE of livestock possible will be described, then the methods currently used, and finally emerging technologies that promise to make GE simpler, more precise, and reliable.

Assisted Reproduction Techniques

Reproduction is the key biological function that is exploited to make a GE animal. Other approaches can be used, however, if the procedures do not target the germ line or the very early embryo then the genetic modification will not be inherited by the offspring and will be lost in the following generation. This is why the transgene has to be introduced either in the gametes or in an embryo at the very early stage of development, ideally when the embryo it is at the one-cell stage. Therefore, mastering assisted reproduction techniques is the key for the successful generation of GE mammals including livestock (Fig. 1).

Male Germ Line

Male germ cells are located in the testis. In the testis, all stages of development are found from spermatogonial stem cells to mature spermatozoa. Spermatogonia can be collected through a biopsy, cultured *in vitro* [11] and when are reintroduced in the testis they can give



Transgenic Livestock Technologies. Figure 1

Summary of Reproductive Technologies (*in counterclockwise*) Ovarian immature oocytes are recovered from ovaries at slaughterhouse, matured in vitro by culturing in a specific medium, fertilized in vitro by co-incubation with capacitated spermatozoa or microinjected with a spermatozoa with a glass micropipette. The resulting zygotes are cultured in vitro until developed to blastocyst stage and finally transferred to the uterus of a synchronized recipient female by a surgical or a nonsurgical procedure depending on the species

rise to functional spermatozoa [12, 13]. During in vitro culture, spermatogonial stem cells can be GE and later transplanted back to the testis to restore spermatogenesis [14]. More easily from sexually mature animals, spermatozoa are recovered from the ejaculate, extended with appropriate diluents, and stored for fresh insemination or cryopreserved for later use. This is a well-established technique in livestock for use in artificial insemination. Spermatogenesis continues after puberty throughout the life of the animals producing an unlimited amount of spermatozoa. In an alternative route, prior to artificial insemination, spermatozoa can be exposed to the transgene in vitro. Spermatozoa can bind DNA and then convey it inside

the oocyte at the time of fertilization [15]. Although the technique is not always reproducible [16], in this simple way some success in making GE livestock has been reported [94–96].

Female Germ Line and the Embryo

In mammals, primordial follicles, the functional unit that contains the female germ cell, exist on the ovaries but only a tiny proportion of the oocytes are effectively generating a progeny while almost all of the remaining are lost during lifetime through atresia. For many years, several investigators isolated and investigated the biology of intact primordial follicles in rodents [17] and finally grew them successfully in culture [18]. This task

proved much more difficult in large animals with bigger ovaries having a lot of connective tissue [19]. Successful development with the birth of live calves, albeit at a low rate, was later obtained by culturing bovine oocytes from more advanced (secondary) follicles [20]. The culture requirements for growing a complex tridimensional structure, such as an ovarian follicle, are so complex that the only progress that has been made is the culture in vivo of thin slices of the ovary following transplantation in humans where this now has a potential application in patients undergoing certain therapies [21].

From the late 1970s, it became clear that also in farm animals it was possible to exploit the oocyte reservoir present in antral follicles from the ovaries of slaughtered animals. Oocytes present in antral follicles have completed the growth phase and they are ready for maturation and fertilization [22, 23]. Initially, successful development was obtained by maturing the oocytes inside the follicles [24]. The subsequent understanding of the crucial role of the surrounding follicular/cumulus cells during the initial phases of maturation [25] lead to the development of a coculture system for in vitro maturation of extrafollicular oocytes obtaining a high proportion of such oocytes developing into live lambs after in vivo transfer [26]. After the successful achievement of in vitro maturation in sheep, the work was expanded investigating in the sheep the effects of gonadotropins [27], and to the cow [28–31] but also to the pig [32] and to the horse [33] and other species as well. Immature oocytes can also be recovered with the ovum pick-up technique in live donor females of different livestock species like cattle [34–36], buffaloes [37], and horses [38]. The development of in vitro maturation is today an irreplaceable, cost-effective and reliable source of mature oocytes for the production of embryos in vitro following in vitro fertilization and the various biotechnological applications connected to it.

Following the seminal work done in laboratory animals [39] in vitro fertilization (IVF) is well developed in farm animal species. Following the birth of the first IVF calf derived from in vivo matured oocytes [40] significant advancements were obtained in IVF when heparin was used as capacitating agent for bull sperm [41]. At about the same time, IVF became a reality in other large species except the horse where the success has

been only exceptional and only one foal has been reported as the result of IVF. The horse eventually benefited of the development of intracytoplasmic sperm injection (ICSI) in humans [42]. The use of ICSI has been very efficient as a way to circumvent failure of in vitro fertilization with a high degree of efficiency [23], not matched in other farm animals like cattle [43, 44], sheep [45], or pigs [46] where the efficiency of ICSI for embryo production is lower as compared to IVF and, therefore, this technique is not used for routine practical application. But this is in fact the way to introduce together with the sperm also the DNA vector required to make the genetic modification [47]. In pigs on the other hand, IVF is characterized by a high incidence of polyspermy that compromises embryonic development. This problem has been overcome in part by improving the quality of in vitro matured oocytes rather than effectively changing IVF conditions.

The immediate goal of increasing the efficiency of in vitro embryo production (IVP) especially in cattle has been the initial driving force of applied research in this field. However, it became soon obvious that merely counting the number of blastocysts was not an accurate measure of the quality of the overall procedure and of the viability of the embryos [48]. For many years, these constraints on the culture of viable cattle embryos in vitro were overcome by using in vivo culture in the oviduct of sheep [7]. In particular, a major concern after in vitro culture of IVF embryos was caused by the description of the so-called Large Offspring Syndrome (LOS), first in sheep and then in cattle [49]. The use of serum supplementation and coculture were later recognized as the primary cause of LOS. Later on, an extensive field study [50] demonstrated that the incidence of LOS was greatly reduced using a SOF (Synthetic Oviductal Fluid) [51]-based formulation or similar formulations of the embryo culture media. These findings clearly indicated that in vitro culture can alter development at very early stage and that the Large Offspring Syndrome is correlated to abnormally advanced embryonic growth and gene expression pattern already at very early stages [52]. Nevertheless the in vitro fertilization and preimplantation embryo culture in vitro allow to generate the number of zygotes required either for DNA microinjection or for virus infection, procedures that would otherwise not be

possible or be very difficult for livestock using *in vivo* produced zygotes from superovulated donors as it is done in laboratory mice.

Nuclear Transfer and Cloning

Somatic cell nuclear transfer, better known as cloning, is a technique that allows the generation of individuals with the same genome; thus technically speaking, cloned animals carry the same genome of the donor cell used for the SCNT process, thus they are like twins between themselves and twins of the animal who donated the somatic cell. The genome comes from the nucleus of a cell that can be taken from an embryo, a fetus, or an animal. The nucleus has to be placed into a mature oocyte, as outlined above whose chromosomes are removed before hand. The nuclear transfer step reconstructs an embryo, with its entire genome originating from the cell's nucleus that will subsequently develop like any other embryo conceived by fertilization albeit at a lower efficiency. Cloning by nuclear transfer that brought to the birth of Dolly the sheep [10] is a milestone achievement that has attracted the attention of the general public toward farm animal embryo technologies. This is one of the best examples where the search of new reproductive techniques in farm animals has also provided basic knowledge not previously obtained in laboratory animals. The fundamental work of cloning by nuclear transfer was done in amphibia [53] in the 1960s, however, significant advancements of the technology were achieved with embryo cloning of farm animals [54] when matured oocytes, instead of zygotes, became the recipients of the donor nuclei. Ten years later, it was again farm animals that, with somatic cell nuclear transfer, gave a major contribution to the basic understanding of cell reprogramming and epigenetic control of mammalian development, opening a new era in cell biology. The success with which different mammals were cloned was directly correlated to the availability of good quality mature oocytes and good quality embryos either after *in vivo* or *in vitro* culture. Then, it was merely a combination of already available technologies of gamete and embryo manipulation that lead to success and, for example, in the horse it was not until such technologies were developed that cloning was achieved in a consistent and reproducible way [55, 56]. In the

last 10 years, the efficiency of somatic cell nuclear transfer as a whole has not improved much and its practical application is limited to the production of animals that have high added value like breeding stock [56–58] or for generating transgenic founder animals when nuclear transfer is combined with genetic engineering of somatic cells [59–62].

Understanding the constraints of nuclear transfer and developing the concept of reprogramming of fully differentiated somatic cells and the derivation of stem cells from cloned embryos [63, 64] for potential biotechnology and therapeutic applications, has fostered the idea of reprogramming differentiated cells *in vitro* directly without the need of the oocyte. A set of four genes has been demonstrated to be at the core of pluripotency and when transfected to somatic cells gave rise to induced pluripotent stem cells (iPS cells) [65]. These cells, of somatic origin, are extraordinary because they carry most of the properties of embryonic stem cells derived from the early embryo and represent a major breakthrough to overcome the limitation of embryonic stem cells derivation in humans and livestock.

Once the molecular basis of pluripotency will be fully understood, it will be a major step forward that is expected to lead to advances in nuclear transfer technology and cloning of farm animals for agricultural and biomedical applications on a larger, solid basis.

Embryonic Stem Cells

Embryonic stem cells (ES cells) are pluripotent cells that originate from the early embryo either from the inner cell mass or from the early epiblast. ES cells are cultured and expanded *in vitro* in undifferentiated conditions, can be GE if necessary, and when reintroduced in the embryos can give rise to any cell type including the germ line, except trophoblast. This is why they are defined as pluripotent and not totipotent cells which is a feature of gametes, zygotes, and early cleavage stages blastomeres only. Today, mouse ES cells technology is the principal route to make GE offspring [8].

The derivation of farm animal ES cells, equivalent to those described for the mouse, has not been reported yet (therefore this route for GE is not available at present in livestock species but it might be so in the near future). However, over the years, from 1981 when mouse ES cells were first discovered [66, 67], many laboratories have

attempted ES cell derivation mainly from cattle, pig, and sheep embryos [9, 68]. The original mouse approach has been extensively used and finally proven, with no doubt, unsuccessful although a few reports have been published on the derivation of so-called ES-like cells [69]. However, the stemness (pluripotency) of these cells appeared to be very limited and most likely they represent trophoblastic cells given their epithelial nature, loss of OCT4 expression, and limited differentiation potential [68–70]. The prominent tendency for neural differentiation often observed when attempting ES cell derivation from large animals is another challenge, but it represents an interesting and robust *in vitro* model for the study of early neurulation events in mammals [64].

In light of all the published reports and observations, it appears that the search for ES cells in farm animals should abandon the original mouse procedure and pay more attention to the recent findings for ES derivation in mouse and humans based on different signaling pathways and/or specific small molecules inhibitors. Advances in mouse and, more recently, in human ES cells culture have demonstrated that a number of different culture conditions can support pluripotency of embryo-derived stem cells. Mouse ES cells can be grown not only in serum-supplemented cultures with LIF and feeders but also in serum-free and feeder-free culture by the addition of BMP molecules. This second method has been developed following the finding that mouse ES cells grown in serum-supplemented medium and LIF can be differentiated in neuroectodermal cells by serum and LIF withdrawal [71]. The role of BMP molecules is to counteract and block the induction of neural differentiation and fix the undifferentiated state in a clever balance between conflicting inductive signaling pathways [71]. Along the same lines, a novel approach for ES derivation in the mouse has been developed on the rationale that the undifferentiated state can be maintained in culture by simply shielding the pluripotent cells of the embryos from endogenous pro-differentiation FGF4 signaling. This method is based on the use of specific inhibitors of the FGF signaling cascade in association with GSK3 inhibitors, the latter acting by improving cell viability. Both mouse and, remarkably for the first time, also rat ES cells have been obtained by inhibition of differentiation signaling introducing the concept that ES cells represent a ground-state pluripotency capable of self-renewing providing that differentiation signals are blocked [72, 73].

Other protocols, based on the stimulation of the nodal-activin signaling pathways, have been shown to maintain the undifferentiated proliferation of human ES cells and mouse epiblast stem cells (EpiSCs). The latter are a type of pluripotent embryo-derived stem cells derived from the egg cylinder as compared to ES cells that derive from inner cell mass or very early epiblast cells. Interestingly, mouse EpiSCs have been shown to be very similar to human ES cells in morphology, growth factors requirement, and gene expression [74, 75] while mouse ES cells differ considerably from human ES in culture requirements for the maintenance of the undifferentiated state, growth rate, and response to inductive signals. Another important difference between mouse ES cells and EpiSCs is the fact that only the former are capable of giving rise to chimeric offspring following blastocyst injection. A recent publication has reported the first success in the derivation of pig EpiSCs [76] and represents a significant step forward toward understanding the requirements for maintaining pluripotency of cultured epiblast cells from farm animal embryos.

In an applied perspective, embryonic stem cells in farm animals are important for several reasons but the most relevant is to provide a method to introduce precise genetic modification into animals by homologous recombination in ES cells [60] followed by blastocyst injection for chimera derivation and breeding, or by somatic cell nuclear transfer. A second important objective is to provide large animal models in which the ES cell technology can be tested for tissue-specific differentiation [77] and cell therapy of various tissues and organs.

Techniques for Genetic Engineering

The meaning commonly given to the term “transgenic” or “GE-” animal, indicate an individual bearing an exogenous fragment of DNA integrated in its genome and transmitting such modification to its descendant through a Mendelian fashion. Therefore, situations like “somatic transgenesis” will not be taken into consideration, where the genetic modification is limited to somatic tissues e.g., in case of gene-therapy protocols applied to adult individuals) or “episomal transgenesis”, where the exogenous DNA is present inside the cell nucleus but as an autonomous

replicating entity, whose stability and replication are separated from those of host's chromosomes.

The natural processes driving the integration of a transgene into the chromosomes of the host cell are mediated by a set of enzymatic function related to the DNA-repair pathways. The enzymes involved in these pathways are responsible of repairing the host's DNA molecule in case of double strand break (DSB). The resolution of a DSB inside the cell nucleus may follow two alternative pathways, catalyzed by different enzymes. The first and more common process involves the Non-Homologous End Joining (NHEJ) pathway, where the free ends of DNA generated by the DSB are relinked together. The exogenous DNA introduced in the nucleus may be recruited during this process and integrated between the endogenous arms of a broken chromosome. The integration of a transgene through the NHEJ pathway causes its random positioning into the genome of the host and the frequent integration of multiple copies of the transgene in the same locus, end by end. Both the random integration site and the generation of a multiple-copy array can influence the expression of the transgene in an unpredictable way. The alternative enzymatic machinery available in the cell for repairing DSBs is the Homologous Recombination (HR) pathway. It relies on the availability of a full copy of the broken molecule to provide a one-strand template to restore the original DNA sequence. This enzymatic activity is principally evident during gametogenesis (when it is involved in crossing over between sister chromosomes) and maintains higher level during early embryonic development compared to somatic adult cells.

A transgene may be integrated by HR if it carries a region of homology with the target region in the host genome with the advantage of driving the insertion of the transgene in a specific locus and as a single copy. Nevertheless, this event is still rare compared to random integration by NHEJ. Even in ES cells, that retains a significant HR activity, the event of a targeted insertion of a transgene by HR is two to four orders of magnitude less frequent compared to its random integration.

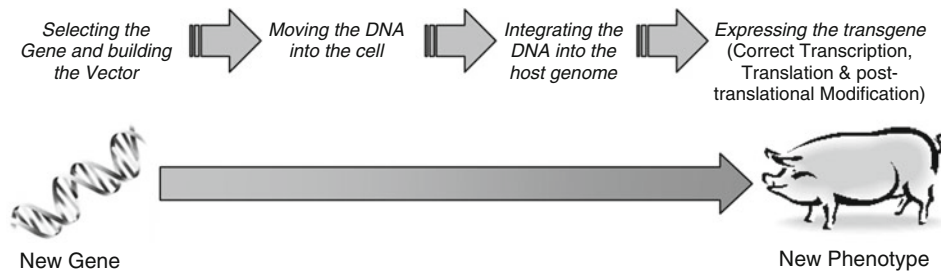
For this reason, the process of gene targeting through natural HR can be performed only on cell cultures, where a large number of integration events can be generated and the cells carrying the desired event can be selected and isolated. The obvious next

step to generate a transgenic animal is the capability of turning these cells into new individuals. In mouse, this has been accomplished by performing gene targeting in ES cells, that are able upon being introduced in a host embryo, to take part to the generation of a chimeric animal including part of its germ line and, as a consequence to pass the acquired mutation to its offspring [78]. As functional pluripotent ES cells have never been obtained from species different than the mouse, the possibility of performing gene targeting has been limited to this species until in 1996 Ian Wilmut and colleagues disclosed the way toward mammalian cloning by SCNT [79].

By then, even though somatic cells retain lower HR activity compared to ES cells, the option of performing gene targeting in these cells and to derive a whole animal from their genome has been demonstrated in swine, ovine, and bovine.

Since the generation of the first transgenic large animals by pronuclear microinjection, including transgenic pigs [3], major advances have been made, mainly by use of assisted reproductive technologies [80] as described earlier. Depending on the purpose of the genetic modification, the approach to transgenesis may aim to the random or the targeted integration of the transgene. In both cases, the obvious but sometime underscored consideration is that the final aim of the process is not only to insert a transgene into the genome, but also its proper *expression*. The hard track toward this is the result of intermediate steps, including (Fig. 2):

1. *Accurate design of the transgene*, involving the selection of the desired gene, its regulatory sequences, and all the additional sequences able to drive its integration and the expression of the desired phenotype in the right tissue at the right time
2. *Introduction of the DNA across the cell membrane* of the embryo or of the cell to be used in nuclear transfer
3. *Integration of the new DNA sequence into the genome*. This step may be mediated only by the endogenous NHEJ or HR pathways but, in some cases, a customized integration of the transgene may be artificially induced by the activity of exogenous recombinases or endonucleases specifically introduced into the cell



Transgenic Livestock Technologies. Figure 2

Steps in obtaining a functional transgenic animal. The definition of transgenic is commonly associated with an organism bearing a new piece of DNA, but the true goal of the procedure is obtaining an animal expressing a new gene according to a desired pattern, that is, showing a new phenotype. Anticipating the final expression pattern of a transgene is still the weaker step in this technology (From [141])

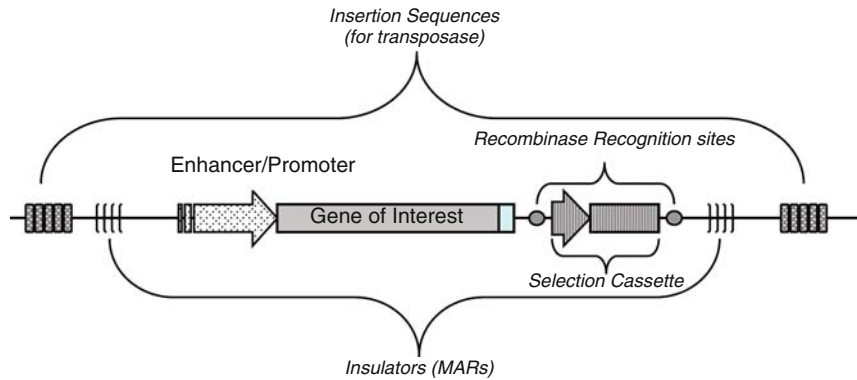
4. *Correct expression of the transgene* is dependent on the good implementation of all of the previously described phases and on a bit of luck. It is usually necessary to producing a large number of different integration events (i.e., embryos or cell clones) to allow the identification and the selection of the best performing, that will be used to generate and expand the desired transgenic line.

During the last years, the toolbox of the genetic engineer has expanded with new instruments allowing a better control on each of the described steps and an overall improvement on the whole procedure. Moreover, some genetic manipulation procedures initially available only for mice, have become concrete even for livestock.

Choosing the Gene and Assembling the Vector

Once the gene to be transferred has been identified, several alternative options are available to generate a transgenic vector (Fig. 3). The coding sequence of the gene itself may be used as cDNA, as genomic sequence (including introns) or as a minigene (a fusion of cDNA and genomic sequences of the same gene). The presence of one or more introns into the transgene, triggering the splicing process, is a major factor improving the transport of mRNA through the nuclear membrane and the effective expression of the exogenous gene [81, 82]. The regulatory sequences to be included into the transgene are chosen according to the need. The original promoter and enhancers may

be conserved to drive the gene according to its original expression pattern, alternatively, heterologous promoters may be used to trigger a different tissue-specific or an ubiquitous expression. In case the gene has to be driven by its original regulatory sequences, the selected genomic fragment should be as large as possible to maximize the chance of a correct expression [83]. Indeed, if a gene has to be transferred from one species to another where the same regulatory signals are recognized, increasing the dimension of the fragment to be transferred allows to including not only the original promoter, but also distal enhancers and chromosomal modifiers [84, 85]. On the other hand, the large size of such DNA molecules influences the choice of the technique required for their introduction into the cell, due to the care required to purify and handle large and fragile DNA molecules. Although microinjection allows the introduction of such big molecules, larger than 100 Kbp, the efficiency of the procedure greatly decreases with the increase of DNA size and a significant part of the resulting transgenic animals integrates only fragmented transgenes, due to shearing of DNA during injection and/or its enzymatic degradation before going into the chromosome. To improve the transfer of large DNA fragment into the genome, specific procedures have been developed, like the ICSI-mediated Gene Transfer (see below). Besides the selection of the gene and its regulatory sequences, additional features may be inserted into the transgenic vector. In case the transgene has to be introduced in cultured cells, it is often required the insertion of a selection



Transgenic Livestock Technologies. Figure 3

Generic structure of a transgenic vector. Beside the expression cassette for the Gene of interest (promoter, CDS and polyA site), additional sequences may be useful to increase the integration frequency of the vector (transposon-derived IS) or to support its regular expression (Insulators). Some features, like drug-resistance genes, are required for isolating cell clones but their persistence in the resulting transgenic organism is undesirable. They can be ablated by means of exogenous recombinases (From [141])

cassette, providing resistance to a specific xenobiotic. As the integration of the vector in the genome of the cell is quite a rare event, the addition of a xenobiotic to the culture medium allows to grow only the transformed cell, that become resistant to it, and to eliminate the rest of the culture.

On the other hand, as the residual presence of this cassette in the transgenic animal may be deleterious or undesired, such cassette may be “floxed” (flanked by lox sites) for a subsequent removal through transient expression of Cre recombinase [86] (see further below).

Random integration of the transgene into the genome is recognized as the main source of variability in the expression of the exogenous genes, due to the influence of flanking genomic regions that may affect the expression pattern of the transgene or “switch it off” by epigenetic silencing. To avoid this behavior, insulator sequences may be inserted to shield the transgene from the influence of the chromosomal location [87–90]. Such insulators, called Matrix Attachment Regions (MARs), bend the transgene in a single transcriptional domain bound by its ends to the nuclear matrix. Additional features of the DNA construct can confer to the vector the behavior of a transposon, being recognized by a transposase that can be cotransferred or transiently co-expressed with the gene during gene transfer. This strategy will be described further below.

Transferring the DNA into the Cell

Pronuclear Microinjection

Pronuclear microinjection is now less commonly used in livestock, having been largely replaced by more efficient and less expensive techniques, after the demonstration of the possibility of cloning mammals by nuclear transfer [10, 91]. Previously, microinjection was considered the most reliable technique to generate transgenic large animals, albeit inefficient and quite expensive in large animals. In vivo production of zygotes provides very viable embryos, but is an expensive approach. The data reported by Krimperfort concerning the generation of the first transgenic calf in 1991, starting from in vitro produced embryos, evidenced the huge effort made in generating this animal.

Two thousand four hundred and seventy (2,470) oocytes were matured and fertilized in vitro to give 1154 zygotes that underwent microinjection; among them, only 129 developed and were transferred in 99 recipient cows. Twenty one (21) calves were born, only two of which were bearing the transgene. One of them died at birth. The other, a male called Herman, grew to adulthood and was mated, because its transgene, coding for human lactoferrin, was designed to be active into the mammary gland of transgenic cows. Sadly, the expression of this gene in the milk of Herman’s daughters was extremely low, making the whole project useless. At the

time, the cost of the whole procedure for producing such transgenic bull was estimated around 500,000 dollars.

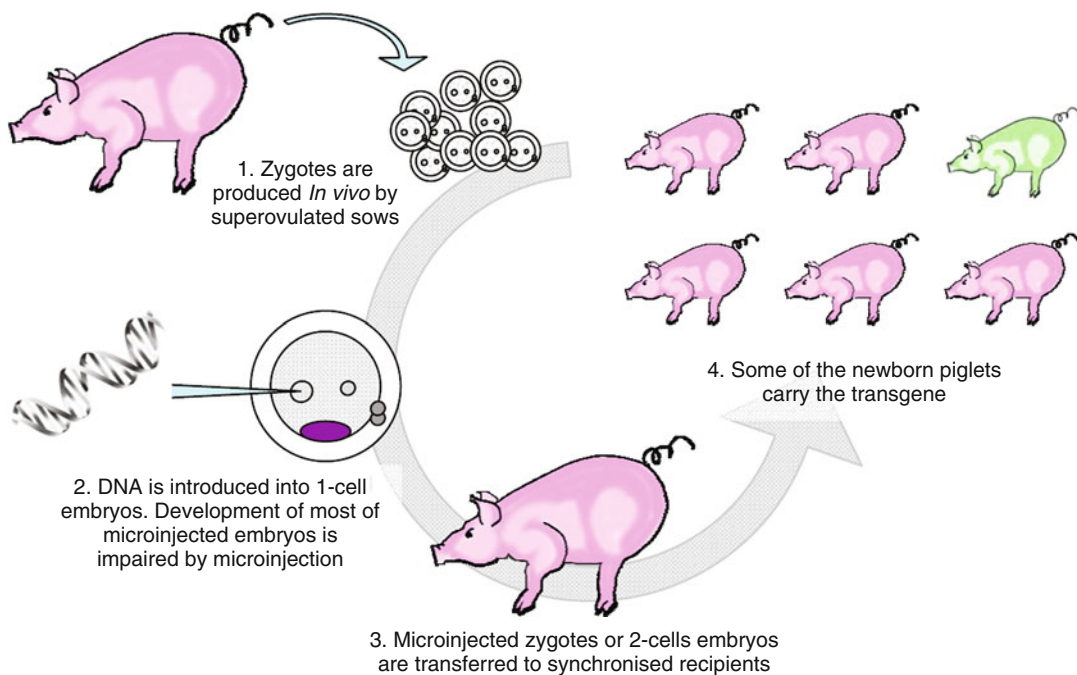
Different factors may affect the expression of microinjected transgenes, considering that they integrate in random sites of the host genome. The disadvantage compared to what happens for gene transfer in cultured cells, in detecting the transgene integration ahead of embryo transfer, although possible, is not practically convenient following pronuclear microinjection, let alone the assessment of its expression. For this reason, most of the animals resulting from microinjection protocols are non-transgenic (Fig. 4).

Last but not least, obtaining embryos bearing targeted transgene integration, until recently, was considered nearly impossible by this approach. DNA integration upon standard microinjection is achieved mostly via NHEJ, resulting in a random positioning

of the transgene into the genome and a consequent unpredictable effect of the flanking regulatory elements on the expression of the transgene itself. Recent advances in enzymatic engineering has demonstrated the possibility to overcome this limit obtaining the targeted integration of a transgene by means of Zinc Finger Nucleases [92] as discussed below.

Sperm-Mediated Gene Transfer (SMGT)

In 1989, Lavitrano and colleagues reported a new approach to generate transgenic mice based on pre-incubation of spermatozoa with exogenous DNA followed by in vitro fertilization [15]. This unconventional approach was accepted with skepticism by many scientists working in the field of transgenesis [16, 93]. In spite of that, other authors reported success in



Transgenic Livestock Technologies. Figure 4

Transgenic production by pronuclear microinjection One-cell embryos are recovered from superovulated females. Such zygotes shows high viability but their production in vivo is somewhat expensive. Embryos are centrifuged to reveal pronuclei and DNA is microinjected into one of them. Microinjection of DNA affects embryo viability. Surviving embryos are transferred in synchronized sows. The proportion of transferred microinjected embryos developing to viable offspring is low and variable, moreover only small fraction of them carries the transgene. The pattern of expression is unpredictable. The overall efficiency of this procedure in terms of living transgenic/microinjected embryos usually ranges from 1% to 4% (From [141])

obtaining transgenic animals by variations of SMGT protocol [94, 95]. After few years, the same technique was adapted to the production of transgenic pigs bearing a hDAF transgene [94–96]. The benefits of this technique, compared to pronuclear microinjection, are low cost and ease of use. Nevertheless, the insertion is still random and the transgene can be rearranged, thus affecting the expression levels. The long-term expression of the transgene remains controversial [97].

ICSI-Mediated Gene Transfer (ICSI-MGT)

ICSI-mediated gene transfer is a technique sharing some analogies with SMGT and microinjection, where both the transgene and the sperm head are introduced with a micropipette into the cytoplasm of the oocyte [47]. In mice, ICSI-MGT is more efficient than standard microinjection and can be particularly effective when a large construct (ranging from 100 kb to more than 0.5–1 Mb, i.e., YAC, BAC, microchromosome) has to be transferred [98]. Together with the use of polyamines to condense DNA, the large size of the micropipette used for injecting the sperm head preserves DNA integrity minimizing mechanical shearing.

Although assisted reproductive technologies (Fig. 1) in pig still suffer some limits, possibly due to specific requirements in embryo culture conditions yet to be defined, some researchers succeeded in producing transgenic pigs by co-incubating sperm with a DNA vector and microinjecting the spermatozoa directly into the ooplasm. After the transfer of 702 embryos into 5 gilts, 2 out of 35 fetuses recovered were transgenic. In vivo production of zygotes provides very viable embryos, but is an expensive approach and requires the use of animals. Nevertheless, this is the preferred source of embryos for this procedure, because microinjection by itself heavily impairs the subsequent embryonic development and decreases the number of newborns following embryo transfer [99].

Viral-Mediated Transgenesis

During their life cycle, retroviruses are able to enter the cytoplasmic membrane of a target cell by binding specific receptor, to reach the nucleus and to retrotranscribe a DNA molecule using their RNA genome

as a template. This DNA molecule will often integrate into the host genome.

Mimicking such behavior, retroviral-derived vectors have been engineered to carry a specific transgene to be integrated into the target genome, but lack the genes required for completing their replicative cycle. Such retroviral vectors, initially developed to target somatic cells for human gene therapy, can also infect preimplantation embryos, providing that the zona pellucida is removed or at least breached, since it is a barrier to the viruses. The retrovirus integrates its complementary DNA into the genome of the embryo. The cell cycle of the infected cell may be susceptible or not to integration depending on the origin of the selected viral vector. For this reason, early trials with retroviral vectors resulted in mosaic animals, due to the failure of these vectors to enter the genome during one-cell stage but only during subsequent embryo divisions. A more recent generation of viral vectors is based on the structure of lentiviruses like HIV. Lentiviral gene transfer is extremely efficient, with 80–100% of the animals being born transgenic after oocyte or embryo, or somatic cell culture infection@ [100]. Lentiviruses have been used in a variety of experiments to transduce cells with various transgenes. These experiments include siRNA knock-down for stem cells and for somatic cells, and nuclear transfer (see below) for generating desired modifications. Lentiviral transgenesis is one of the main techniques currently proposed for somatic gene therapy. Such approach is associated with known risks and observed limits. During clinical trials, retroviral transgenesis has been associated with oncogene activation by insertional mutagenesis [101, 102]. To reduce this risk, the interest of the researchers is moving toward replication-deficient vectors [103]. An additional concern about the use of lentiviral vectors is their possible recombination with latent wild retrovirus to generate unpredictable infectious or mobile particles. Among virus-derived vectors, lentiviral (LV) have the property of infecting cells both during replication and in quiescent phase. During lentiviral-mediated transgenesis, the use of some drugs like cytokines or proteasome inhibitors can increase LV gene transfer [60, 104]. Santoni de Sio and colleagues [104] have shown that human hematopoietic stem cells (HSCs) can be transduced to high efficiency by a short exposure to LVs in the presence of SCF, TPO, IL-6, and Flt3L. Moreover, it was shown that the proteasome restricts LV transduction in HSCs and that using the

reversible peptide–aldehyde proteasome inhibitor MG132 and the peptide-boronate inhibitor PS-341 during the LV-GFP transduction period, there is a substantial drug-dose-dependent increase in the frequency of transgene expressing cells and in their mean fluorescence intensity [104].

The use of lentiviral vectors for transgenesis results in a significant degree of mosaicism among transgenic newborns, due to the possible multiplicity of integration events during the first cleavages of the embryo [105]. Even though lentiviral vectors enter as a single copy into their insertion site, their integration mechanism is very efficient, frequently resulting in multiple copies integrated in different part of the genome and in a complex pattern of transmission to the offspring [105]. Due to their random positioning into the chromosomes, lentiviral vectors may still undergo epigenetic silencing by methylation [106]. Another class of effective vectors for transgenesis was derived from the adeno-associated virus (AAV). AAV-derived vectors have the advantage to cross the cell membrane delivering single stranded DNA molecules straight to the nucleus. Some adeno-associated viruses are peculiar in their behavior of targeting specific integration site in the host genome [107]. Adenoviral ssDNA shows a high effectiveness in integrating into the host genome through homologous recombination, even in somatic cells [108, 109]. Recently, Rogers et al. produced a CFTR-null pig using AAV-mediated gene targeting and SCNT [110, 111]. These authors generated a pig with both null (knock-out) and $\Delta F508$ (knock-in) modifications. Gene targeting using the AAV approach has resulted in a very efficient strategy for obtaining knock-out of the CFTR gene that is not expressed in fibroblasts.

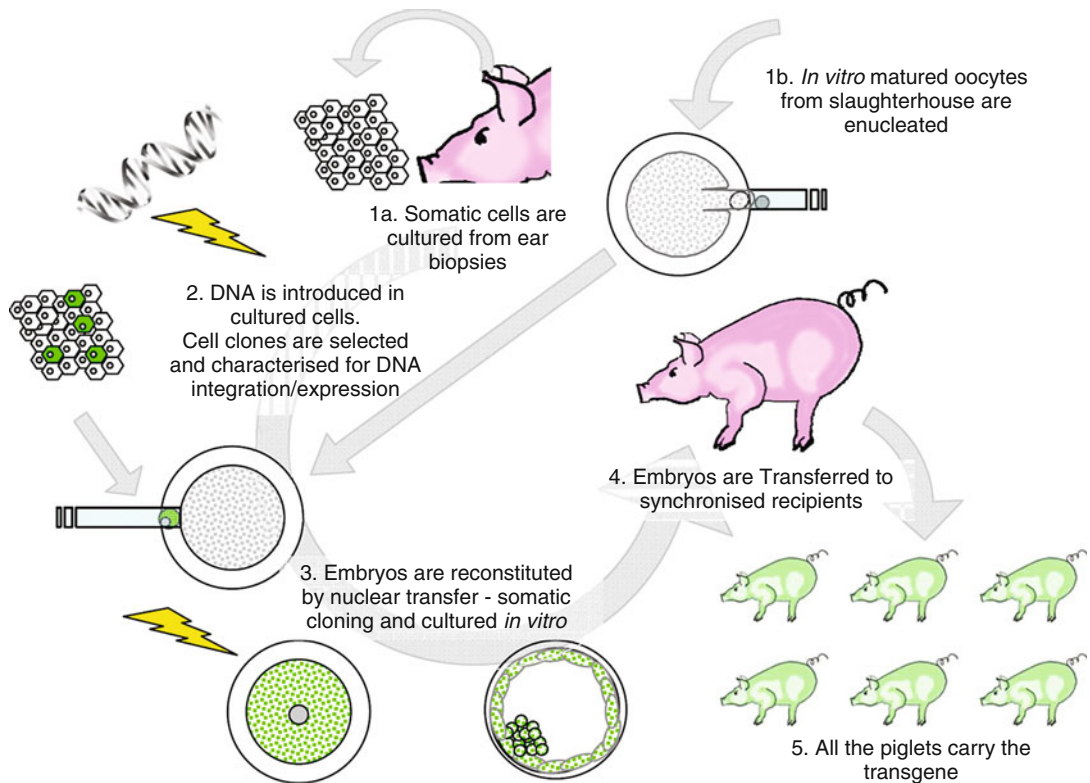
Somatic Cell Nuclear Transfer

Somatic cell nuclear transfer (SCNT) has become the leading tool for generating animals from GE somatic cells (Fig. 5). SCNT works better in pigs than in other large animals [112, 113]. A recent innovation to make the technique user-friendly is the zona-free system. After zona removal, enucleation can be performed with a micropipette, for subsequent zona-free fusion, activation, and culture [113, 114], or by cutting the cytoplasts for handmade cloning [115, 116]. Major limitations are represented by the lack of embryonic stem cell (ESC)

technologies. Somatic cells are currently being used in the procedures, but they have a limited life span, thus restricting the time the cells can be cultured in vitro for genetic engineering. Fibroblasts (both from fetal and adult origin) have a fine life span and can undergo a maximum of 40–50 population doubling in vitro while maintaining a normal karyotype. Such number of population doublings allow only for a round of cell transfection followed by drug selection to identify a clonal cell population with the desired modification. In fact by the time this has been achieved, the cells have spent most of their proliferative capacity and have to be used for SCNT before they reach complete senescence. So, the use of fibroblasts limits this approach to one modification in vitro at the time since they would not survive a second round of transfection. After SCNT, the fibroblasts are rejuvenated and have fully restored their proliferative capacity. At this point, cloned fetuses are recovered to establish a new fibroblast cell line that can be then subjected to a new round of genetic modification. This process can be repeated several times with apparently no adverse effects [59]; however it requires time and animals. Relating to this topic, recent reports describe the derivation of induced pluripotent stem cells (iPSC) from pig fibroblasts [117–119], however the objective is still far since culture conditions to maintain livestock embryonic stem cells are not known (further discussed below). The opportunity of performing the whole genetic engineering on cell cultured in vitro and finally transferring the manipulated genome to the enucleated oocytes allows selection for gene targeting events, both for *knocking out* an unwanted gene or for *knocking in* a transgene in a predetermined reliable position, guaranteeing its best expression.

A similar model, implementing the 1,3 galactosyltransferase knockout in pigs was developed by Takahagi and colleagues that knocked out the first allele of this gene in porcine fetal fibroblast already expressing hDAF and N-AcetylglucosaminylTransferase -III (GNT-III) [120]. In a further step, the authors selected cultures for spontaneous null mutations of the second allele occurring in vitro, and used these cells for nuclear transfer. In this way, the resulting cloned pigs were transgenic for hDAF and GNT-III on a 1,3 galactosyltransferase null background [121].

SCNT technology overcomes many limitations of previous procedures, making possible the gene targeting



Transgenic Livestock Technologies. Figure 5

Transgenic production by SCNT. The transgene is transferred into the genome of cultured fibroblasts. Transgenic cell clones are isolated and characterized. This first step is relatively inexpensive. A more accurate prediction of the transgene expression is possible. Next, embryos are reconstituted by SCNT, cultured and transferred in synchronized sows. Although the viability of cloned embryos is variable but usually poor, all of the resulting newborns are transgenic (From [141])

approach to livestock genome and providing a less expensive way to perform genetic engineering. Indeed, most of the cost and the efforts related to transgenic technology applied to large animal are derived from housing animals, preparation for embryo transfer up to the birth and weaning of the newborn animals.

The opportunity of using cell clones fully characterized in term of transgene integration and expression, guaranteeing 100% of transgenics among born animals, greatly increases the efficiency of the protocol and drops the costs when an ubiquitous or inducible promoter is used and the expression is maintained in subsequent generations through germ line development (Fig. 6) [112]. On the other hand, SCNT is not the final solution for every problem, as most of the matter concerning the proper activity of the transgenes, with regard to its temporal and tissue-specific

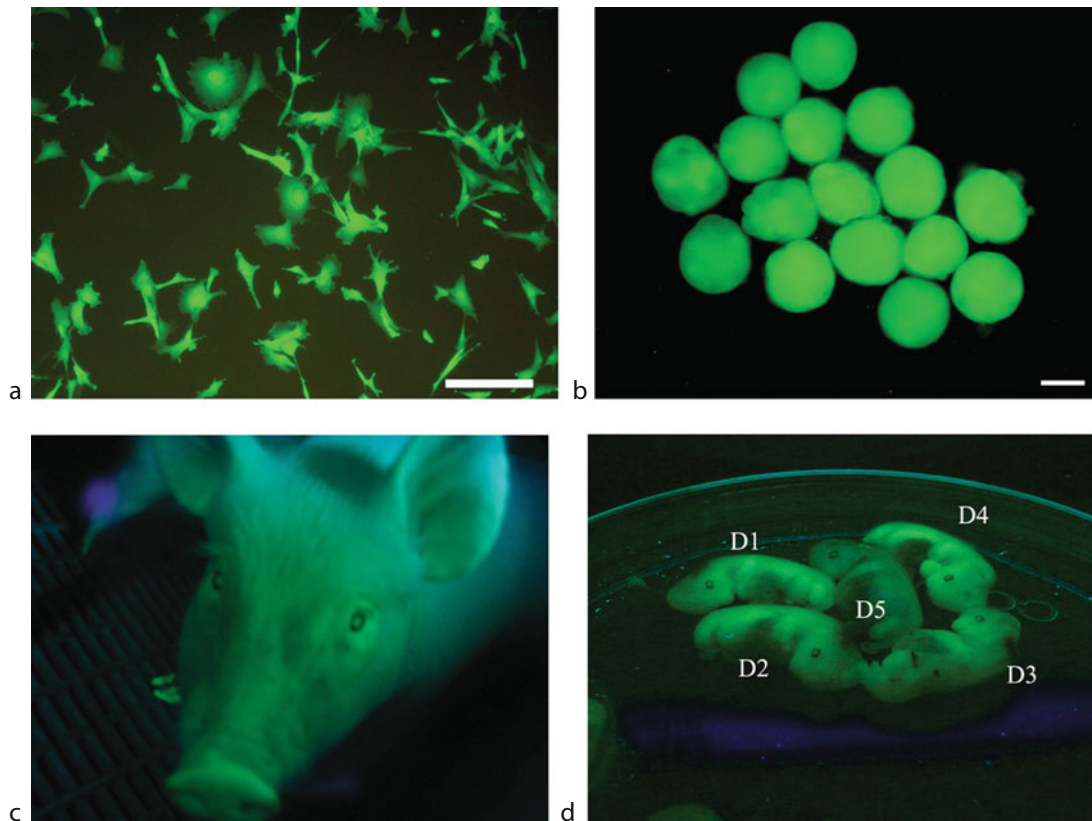
expression, still needs to be completed on the whole animal. Possible improvements to overcome these problems are offered by recent advances in genetic engineering.

Emerging Technologies

Several new technologies becoming available may be of great benefit in the future to make GE large animals.

Induced Pluripotent Stem Cells (iPS Cells)

In 2006, a breakthrough study [65] demonstrated that viral transduction of a handful of genes (Oct4, Sox2, Klf4, and c-Myc) can reprogram mouse embryonic fibroblasts into ES cell-like cells which carry all the molecular features of true embryo-derived ES cells



Transgenic Livestock Technologies. Figure 6

Successful selection of GFP transgenic cells and transmission through the germline of the transgene to the next generation. Clonal fibroblasts selected for high expression of GFP (a), GFP expression in embryos after SCNT (b) and in the resulting animal (c). Fetuses obtained after breeding the GFP boar to a wild type sow (d). Fetus D1, D2, D3, D4 are transgenic and express the same high level of GFP of the original boar, D5 is a wild type fetus negative for GFP (From [141])

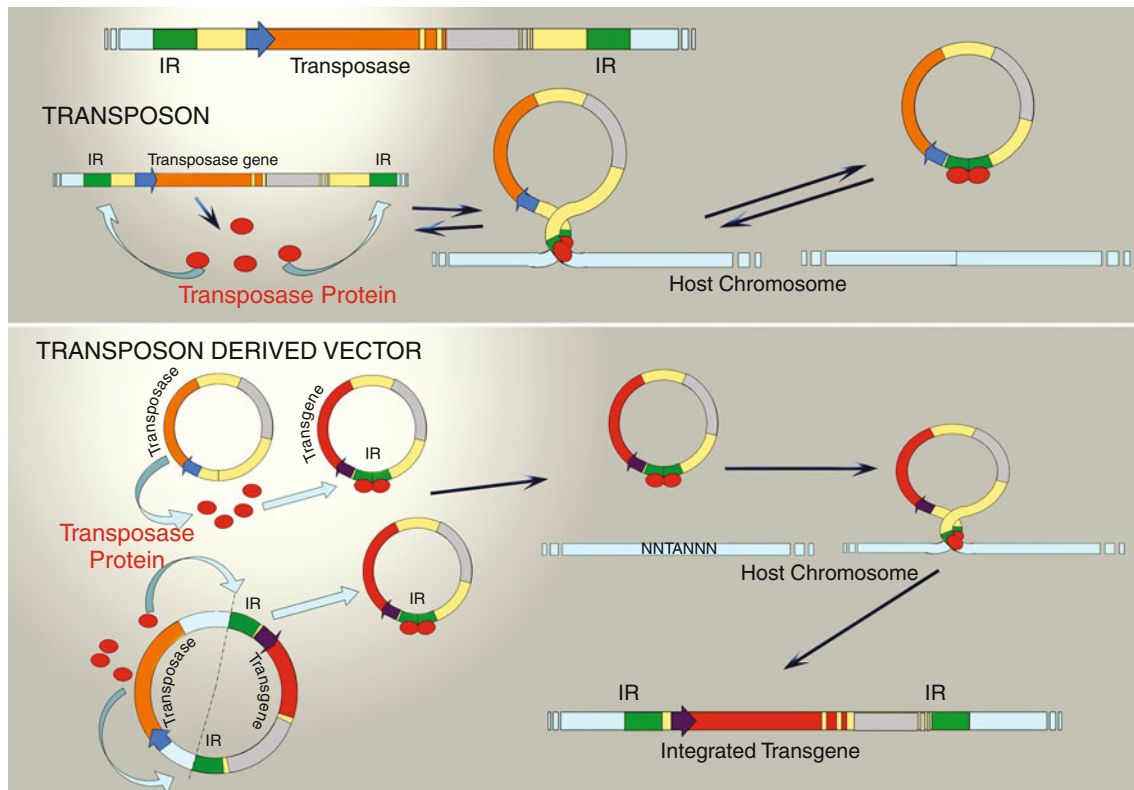
including the ability to give rise to germ line chimeras. The following year the generation of human iPS cells was achieved [122, 123] using slightly modified transduction methods and set of genes. Since then, several groups have contributed to the field with new combinations of reprogramming genes, cell types, and viral and nonviral delivery systems. More recently, the addition of small molecules acting on chromatin structure, such as valproic acid, a histone deacetylase inhibitor has allowed further progress by showing that only Oct4 and Sox2 are required for reprogramming of human fibroblasts [124]. Several other small molecules acting on chromatin or on specific signaling pathways have been investigated to improve the efficiency and safety of iPS cells technology (see as review [125]). In

large animals, attempts to derive iPS cells have been made resulting in a few reports about induced reprogramming of pig fibroblasts [117–119]. Although promising data have been presented, the evidence of full reprogramming, measured by sustained activation of endogenous genes and silencing of reprogramming genes, has not been achieved. Most likely, the lack of robust procedures for the establishment embryo-derived ES cells in large animals represents a major limit also for the development of the iPS cells technology. More detailed knowledge on the role of pluripotency genes in the early embryo is needed to advance the field of induced pluripotency not only in the mouse [126] but also in large animal species.

Enzymatic Engineering

Transposons Transposons, called also “jumping genes” are mobile genetic elements; class II transposons are small segments of DNA able to move across the genome of a cell from one region to another, by means of the action of enzymes (transposase) encoded within the transposon itself or supplied in trans by another

source (Fig. 7). Transposons have been found in many living organism, from bacteria to plants and animals. The simplest autonomous replicating transposons in vertebrates, like those belonging to TC1/Mariner class, are composed by two inverted terminal repeat flanking a sequence coding for a specific transposase. Although the movement of transposons across different



Transgenic Livestock Technologies. Figure 7

Integrative vectors based on Transposon signals *Top*: Natural occurring DNA-transposons consists of a common minimal structure represented by a gene coding for a specific transposase (TP) and two terminal inverted repeats (IRs). Following the expression of the gene, its product is able to bind to the inverted repeats, inducing the circularization and the excision of the complete DNA segment surrounded by IRs. Depending on the type of transposon, excision can leave small footprint in the chromosome or restore the exact original sequence. Following excision, the circular transposon, still bound to transposase, can integrate in a new site of the genome. In presence of transposase, both the excision and the integration of the transposon are catalyzed. Some transposons, like PiggyBac, prefers transcriptionally active genomic region for their integration. *Bottom*: A Transposon-derived vector is represented by a transgene flanked by a pair of IRs, but lacking the transposase CDS between them. The integration of this vector can be obtained with high efficiency by expressing the specific transposase *in trans* – on a different vector (**a**) – or *in cis* – on the same vector but by a cassette located outside the IRs-delimited portion (**b**). The transiently expressed exogenous transposase promotes the integrative cycle of the vector but, due to the absence of a stable source of protein, the integration of the vector is irreversible (From [141])

organism does not seem to be a common capability, phylogenetic studies strongly suggest that “jumps” of transposons across species have happened during evolution [127]. Modified transposons, like “Sleeping Beauty” and “PiggyBac” have been largely used for precise and efficient delivery of DNA expression cassettes in vertebrate cells. Sleeping Beauty (SB) belongs to the TC1/Mariner class of transposons and these transposases require a TA dinucleotide base pair as integration site, a sequence that is duplicated during the integration process. The SB transposon system consists of two components: (1) a defective transposon, made up by gene of interest flanked by inverted repeats (IRs) but lacking the transposase gene, and (2) a source of transposase. During transposition, the SB transposase recognizes the ends of the IRs and excises the transposon from the delivered plasmid DNA inserting it in to another DNA site.

In a recent study, it was shown that co-transfection of PEGE cells with Sleeping Beauty (SB), Passport (PP) Tol2 and PiggyBac (PB), with their corresponding transposase expression constructs, resulted respectively in 13.5-, 5-, 21-, and 28-fold increases over transfection without transposase [128]. In addition to increasing the efficiency of integration, transposase-mediated transgenesis precisely integrates a single copy of the transposon into one or more locations in the genome, avoiding transgene concatemerization that can cause shutdown of gene expression.

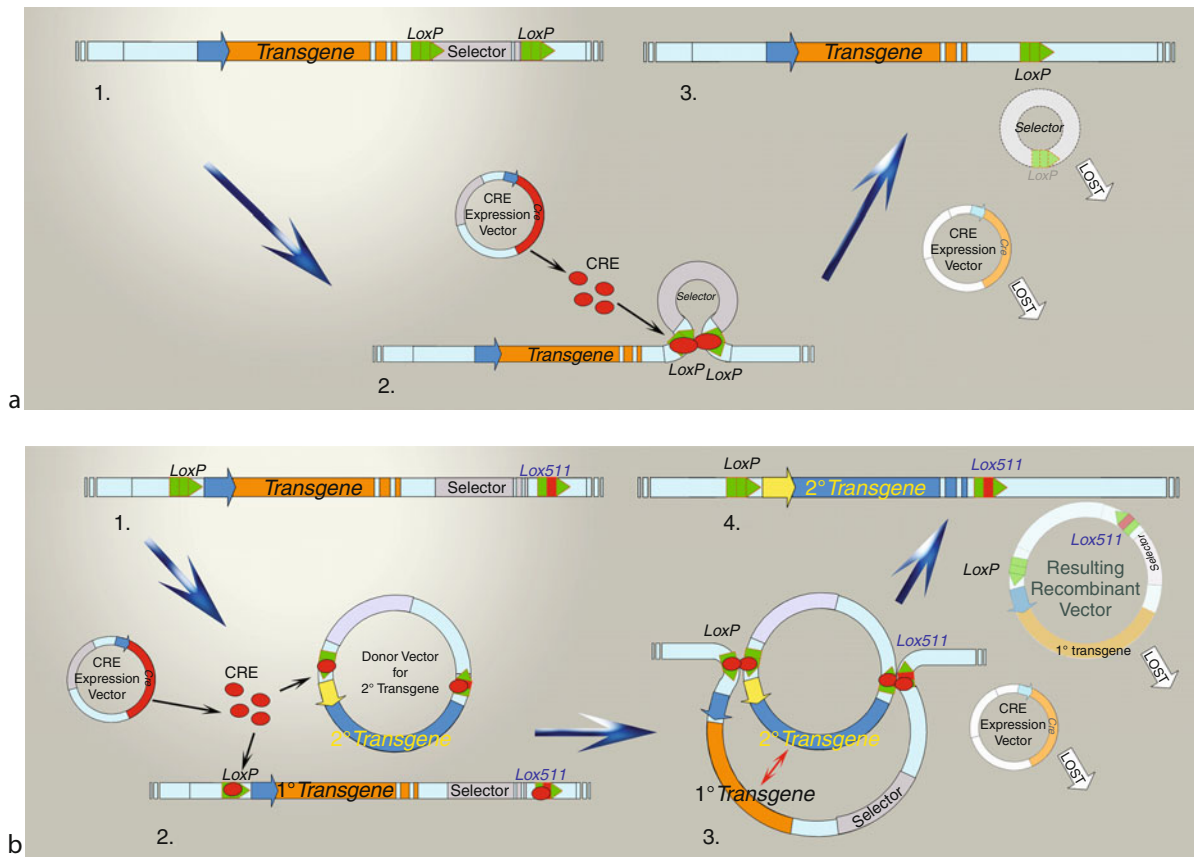
Cre/LoxP Recombinases The integration site of a transgene strongly influences its expression pattern. To overcome the gamble of a random integration, it is possible to target the insertion of the transgene to a transcriptionally active location of the genome, avoiding the risks of both silencing the transgene and disrupting an endogenous gene (insertional mutagenesis).

One possible strategy to obtain this goal is the use of phage recombinases like Cre or FLP that catalyze a conservative DNA recombination event between two short recombinase recognition sites (RRS), loxP and FRT (Fig. 8), respectively. Such enzymes, used by bacteriophages during their infection cycle, allows the excision or inversion of the DNA between two RRSs, depending on their orientation [128]. The artificial modification of the sequences

of the parental RRSs has allowed to develop the so-called Recombinase-Mediated Cassette exchange (RMCE), a protocol to modify a specific locus in the genome after an initial “tagging” by the introduction of a pair of incompatible RRSs [129]. In this way, (1) a cell line is modified by inserting a RRS-flanked reporter gene in different random position, (2) the deriving clones are screened to select the one presenting the best integration site, and (3) the gene of interest can be exchanged with the reporter to assure its proper expression. A predictable modification of the genome may be realized by driving the process of DNA repair through homologous recombination (gene targeting). This process, largely used in mouse ES cells, allows the interruption of endogenous sequences (knock-out) or the insertion of new genes (knock-in) in a specific locus [130, 131]. Unfortunately, the HR pathway is much less efficient in somatic cells like fibroblasts, making this procedure much harder, in particular for genes that are inactive in the target cell type not allowing the use of gene-trap (promoterless) approaches.

Zinc Finger Nucleases (ZFNs) Zinc finger nucleases (ZFNs) show promise in improving the efficiency of gene targeting by introducing DNA double-strand breaks in target genes, which then stimulate the cell’s endogenous HR machinery. Zinc finger nucleases are hybrid proteins containing an array of zinc-finger DNA-binding domain and a FokI endonuclease domain (Fig. 9). The DNA zinc-finger domains are designed to recognize a specific sequence, inducing a double stranded break in the target site. Such break promotes a local DNA repair activity that is efficiently accomplished by HR if a template with homologous sequence is provided. Many studies have been developed in human and mouse cells [60, 132]. A recent paper demonstrates the possibility of generating transgenic rats by pronuclear microinjection of specific ZFN expression vectors together with a gene targeting construct [92].

An additional requirement in developing transgenic livestock is the need of expressing multiple transgenes in a coordinate pattern in the same animal. As previously noted, protocol bringing to random integration of multiple transgenes, often result in similarly random level of expression, with some



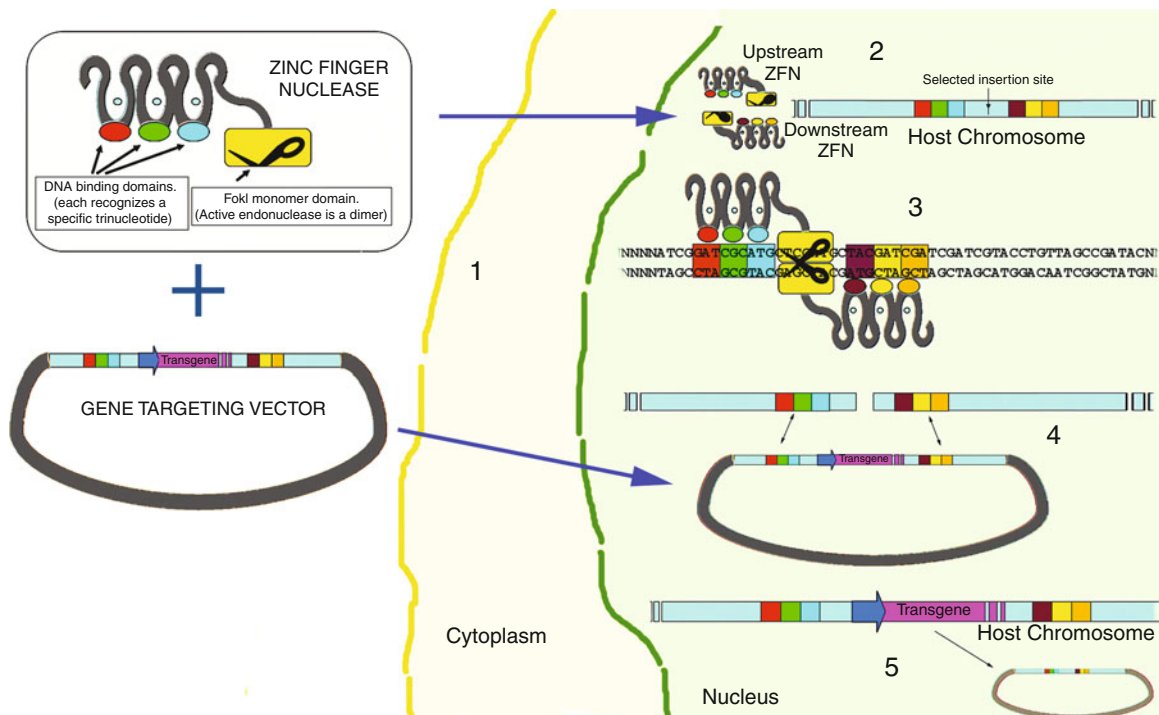
Transgenic Livestock Technologies. Figure 8

Targeted rearrangement of an integrated transgene by Cre recombinase. **(a)** Homologous Lox sites in the same orientation allows the excision of the enclosed region when Cre is expressed **(b)** Heterologous Lox sites do not allow excision but promote RMCE (Recombinase Mediated Cassette Exchange) if a vector bearing a Lox-flanked transgene is introduced together with CRE

transgenes working and some other not. A strategy to speed multiple transgene integration is represented by recent adaptation of the 2A system from foot and mouth disease virus (FMDV) to mammalian transgenic technology [133, 134]. In this system, the open reading frame (ORF) consists of multiple individual cDNAs separated by sequences encoding 2A and furin cleavage sites. A single complex mRNA is produced and translated into a single polypeptide that is cleaved into individual exogenous proteins at the 2A sites.

siRNA In cells transfected with siRNA vectors, targeted mRNAs are degraded by endonuclease activity and the amount of protein translated may be reduced

by over 95%, thus resulting in a significant knock-down and is an alternative approach to achieving more complex and difficult knock-outs (KO). This technique is particularly useful when more than one copy of the endogenous gene is present and the usual KO approach is not feasible. This is indeed the case of pathogens like porcine endogenous retrovirus (PERV) [135–137]; for example, “knock-down” of PERV expression has been shown in transgenic pigs expressing siRNA corresponding to the viral pol2 sequence [138]. This approach could even be of interest in generating transgenic animals resistant to specific viral pathogens, by steady expression of virus-inhibiting siRNA molecules in their cells.



Transgenic Livestock Technologies. Figure 9

Homologous Recombination via Zinc Finger Nucleases (ZFNs) 1. A gene targeting DNA vector is prepared, where the transgene is flanked by DNA sequences homologous to the target locus. In different vectors, two ZFNs are designed, each coding for a specific DNA binding region and a FokI endonuclease monomeric domain. The two sequences recognized by the ZFNs are positioned upstream and downstream the selected insertion site in the target locus. The ZFN is introduced into the cell by microinjection, electroporation or transfection of a DNA expression vector or of a mRNA transcript. 2. The mRNAs are translated in two Zinc Finger Proteins. Each "finger" motif of the ZFN recognizes a sequence of three nucleotide (in this example, three "fingers" recognize a 9 bp target). A single ZFN binding to the DNA does not produce any effect. 3. When both the ZFN bind the DNA at the right distance, FokI domains can dimerize and produce a Double Strand Break in the target site. 4. DNA repair on the DSB can proceed through the NHEJ pathway, but, due to the availability of homologous sequences provided by the targeting vector, can also follow the route of Homologous Recombination. 5. By Homologous Recombination, the transgene is introduced into the target site. The endogenous allele and the vector backbone sequences are lost (From [141])

Future Directions

The GE of livestock is slowly making progress toward possible applications both in science and industry. SCNT has been the major advancement in this field for the last 25 years. Nevertheless, many are the issues that need to be addressed to make GE of livestock robust, reproducible, and affordable. Assisted reproduction techniques are continuously improved and refined, however, the availability of true embryonic stem cells is still the one of the limiting factors for precise GE and the recent report of

alternative sources like iPS cells have yet to prove their value for this purpose and it might take a long time before they become a reality. The designing of more effective DNA delivery vectors relies on the sequence of the genome. In livestock, this is lagging behind that of the human or mouse, therefore, it is not always possible to translate discoveries made in those species to livestock unless genomic information are generated by homology and partial sequencing. The ability to target the insertion of the transgene into a locus on a chromosome to assure

its expression also through generations would be a significant advancement since today few of the GE livestock expresses the inserted transgene at the level desired and the expression is often not stable.

The ability to target specific genes either by knockout through HR or knockdown by siRNA, will make it possible to develop functional genomics (i.e., understanding the function of a particular gene) in mammalian species closer to the human other than the mouse and to generate large animal models of human diseases, especially for those diseases that the mouse model has failed to reproduce the human phenotype [110, 139]. Animal models are required both to understand the pathogenesis of a disease, to develop and test possible new therapeutic approaches to it. Biotechnological application of GE animals again for biomedical application include the field of xenotransplantation, i.e., the possibility to transplant tissues and organs from one species (usually the pig) to humans [140, 141]. This requires multiple GE to make pig organs accepted by the human immune system. Another biotech application is the GE of livestock to express antibodies or proteins that have a therapeutic use in human cancer or other diseases [59] in sufficient quantities and affordable costs that would not otherwise be possible by other means.

Although the biomedical field is the more advanced and more receptive for this new technology, there are also potential applications in agriculture for breeding and selection of livestock. Proof of principle have been already obtained for increasing yield in cheese production [62] or better quality pork meat [142] as well as for disease resistance [61, 138].

GE of livestock is an expanding area of research with great potential for applications in the biomedical, biotechnology, and agricultural fields. This technology provides a tremendous potential to select and breed animals with specific genetic makeup much faster than by conventional breeding schemes; however, this will be possible when full understanding of the underlying biological mechanisms will be known.

Acknowledgments

Grant support was received from EU Xenome Project (LSHB-CT-2006-037377), EU PluriSys (n°223485), EU ESNATS (n° 201619), Regione Lombardia and Cariplo Foundation (NOBEL).

Bibliography

- Gordon JW, Scangos GA, Plotkin DJ, Barbosa JA, Ruddle FH (1980) Genetic transformation of mouse embryos by microinjection of purified DNA. *Proc Natl Acad Sci USA* 77:7380–7384
- Palmiter RD, Brinster RL, Hammer RE, Trumbauer ME, Rosenfeld MG, Birnberg NC, Evans RM (1982) Dramatic growth of mice that develop from eggs microinjected with metallothionein-growth hormone fusion genes. *Nature* 300:611–615
- Hammer RE, Pursel VG, Rexroad CE Jr, Wall RJ, Bolt DJ, Ebert KM, Palmiter RD, Brinster RL (1985) Production of transgenic rabbits, sheep and pigs by microinjection. *Nature* 315:680–683
- Simons PJ, Wilmut IA, Clark J, Archibald AL, Bishop JO, Lathe R (1988) Gene transfer into sheep. *Biotechnology* 6:179–183
- Wall RJ, Pursel VG, Shamay A, McKnight RA, Pittius CW, Hennighausen L (1991) High-level synthesis of a heterologous milk protein in the mammary glands of transgenic swine. *Proc Natl Acad Sci USA* 88:1696–1700
- Krimpenfort P, Rademakers A, Eyestone W, van der Schans A, van den Broek S, Kooiman P, Kootwijk E, Platenburg G, Pieper F, Strijker R (1991) Generation of transgenic dairy cattle using 'in vitro' embryo production. *Bio/technology (Nature Publishing Company)* 9:844–847
- Galli C, Lazzari G (1996) Practical aspects of IVM/IVF in cattle. *Anim Reprod Sci* 42:371–379
- Capecchi MR (1989) Altering the genome by homologous recombination. *Science* 244:1288–1292
- Galli C, Lazzari G, Flechon JE, Moor RM (1994) Embryonic stem cells in farm animals. *Zygote* 2:385–389
- Wilmut I, Schnieke AE, McWhir J, Kind AJ, Campbell KH (1997) Viable offspring derived from fetal and adult mammalian cells. *Nature* 385:810–813
- Brinster RL (2007) Male germline stem cells: from mice to men. *Science* 316:404–405
- Brinster RL, Zimmermann JW (1994) Spermatogenesis following male germ-cell transplantation. *Proc Natl Acad Sci USA* 91:11298–11302
- Dobranski I (2008) Male germ cell transplantation. *Reprod Domest Anim* 43(Suppl 2):288–294
- Brinster RL (2002) Germline stem cell transplantation and transgenesis. *Science* 296:2174–2176
- Lavitrano M, Camaioni A, Fazio VM, Dolci S, Farace MG, Spadafora C (1989) Sperm cells as vectors for introducing foreign DNA into eggs: genetic transformation of mice. *Cell* 57:717–723
- Brinster RL, Sandgren EP, Behringer RR, Palmiter RD (1989) No simple solution for making transgenic mice. *Cell* 59:239–241
- Roy SK, Greenwald GS (1985) An enzymatic method for dissociation of intact follicles from the hamster ovary: histological and quantitative aspects. *Biol Reprod* 32:203–215
- Eppig JJ, O'Brien MJ (1996) Development in vitro of mouse oocytes from primordial follicles. *Biol Reprod* 54:197–207
- Lazzari G, Galli C, Moor RM (1994) Functional changes in the somatic and germinal compartments during follicle growth in pigs. *Anim Reprod Sci* 35:119–130

20. Miyano T, Manabe N (2007) Oocyte growth and acquisition of meiotic competence. *Soc Reprod Fertil Suppl* 63:531–538
21. Meirou D, Levron J, Eldar-Geva T, Hardan I, Fridman E, Yemini Z, Dor J (2007) Monitoring the ovaries after autotransplantation of cryopreserved ovarian tissue: endocrine studies, in vitro fertilization cycles, and live birth. *Fertil Steril* 87:418 e417–418 e415
22. Galli C, Moor RM (1991) Development of immature bovine oocytes into viable embryos in vitro. *Bull 'Association Anat* 228:67–71
23. Galli C, Colleoni S, Duchi R, Lagutina I, Lazzari G (2007) Developmental competence of equine oocytes and embryos obtained by in vitro procedures ranging from in vitro maturation and ICSI to embryo culture, cryopreservation and somatic cell nuclear transfer. *Anim Reprod Sci* 98:39–55
24. Moor RM, Trounson AO (1977) Hormonal and follicular factors affecting maturation of sheep oocytes in vitro and their subsequent developmental capacity. *J Reprod Fertil* 49:101–109
25. Moor RM, Osborn JC, Cran DG, Walters DE (1981) Selective effect of gonadotrophins on cell coupling, nuclear maturation and protein synthesis in mammalian oocytes. *J Embryol Exp Morphol* 61:347–365
26. Staigmiller R, Moor R (1984) Effect of follicle cells on the maturation and developmental competence of ovine oocytes matured outside the follicle. *Gamete Res* 9:221–229
27. Galli C, Moor RM (1991) Gonadotrophin requirements for the in vitro maturation of sheep oocytes and their subsequent embryonic development. *Theriogenology* 35:1083–1093
28. Carolan C, Monaghan P, Gallagher M, Gordon I (1994) Effect of recovery method on yield of bovine oocytes per ovary and their developmental competence after maturation, fertilization and culture in vitro. *Theriogenology* 41:1061–1068
29. Lu K, Polge C. A summary of two-year's results in large scale in vitro bovine embryo production. *Proc 12th Inter. Cong. Animal Reproduction, The Hague* 1992:1315–1317
30. Goto K, Kajihara Y, Kosaka S, Koba M, Nakanishi Y, Ogawa K (1988) Pregnancies after co-culture of cumulus cells with bovine embryos derived from in-vitro fertilization of in-vitro matured follicular oocytes. *J Reprod Fertil* 83:753–758
31. Xu KP, Greve T, Callesen H, Hyttel P (1987) Pregnancy resulting from cattle oocytes matured and fertilized in vitro. *J Reprod Fertil* 81:501–504
32. Moor RM, Mattioli M, Ding J, Nagai T (1990) Maturation of pig oocytes in vivo and in vitro. *J Reprod Fertil Suppl* 40:197–210
33. Zhang JJ, Boyle MS, Allen WR, Galli C (1989) Recent studies on in vivo fertilization of in vitro matured horse oocytes. *Equine Vet J* 8:101–104
34. Galli C, Crotti G, Notari C, Turini P, Duchi R, Lazzari G (2001) Embryo production by ovum pick up from live donors. *Theriogenology* 55:1341–1357
35. Pieterse MC, Vos PL, Kruip TA, Wurth YA, van Beneden TH, Willemse AH, Taverne MA (1991) Transvaginal ultrasound guided follicular aspiration of bovine oocytes. *Theriogenology* 35:857–862
36. Hasler JF, Henderson WB, Hurtgen PJ, Jin ZQ, McCauley AD, Mower SA, Neely B, Shuey LS, Stokes JE, Trimmer SA (1995) Production, freezing and transfer of bovine IVF embryos and subsequent calving results. *Theriogenology* 43:141–152
37. Galli C, Duchi R, Crotti G, Lazzari G (1998) Embryo production by Ovum Pick Up in Water Buffalo. *Theriogenology* 49:400
38. Galli C, Crotti G, Turini P, Duchi R, Mari G, Zavaglia G, Duchamp G, Daels P, Lazzari G (2002) Frozen-thawed embryos produced by Ovum Pick Up of immature oocytes and ICSI are capable to establish pregnancies in the horse. *Theriogenology* 58:705–708
39. Yanagimachi R, Chang MC (1964) In Vitro Fertilization of Golden Hamster Ova. *J Exp Zool* 156:361–375
40. Brackett BG, Bousquet D, Boice ML, Donawick WJ, Evans JF, Dressel MA (1982) Normal development following in vitro fertilization in the cow. *Biol Reprod* 27:147–158
41. Parrish JJ, Susko-Parrish JL, Leibfried-Rutledge ML, Critser ES, Eyestone WH, First NL (1986) Bovine in vitro fertilization with frozen-thawed semen. *Theriogenology* 25:591–600
42. Palermo G, Joris H, Devroey P, Van Steirteghem AC (1992) Pregnancies after intracytoplasmic injection of single spermatozoon into an oocyte. *Lancet* 340:17–18
43. Goto K, Yanagita K (1995) Normality of calves obtained by intracytoplasmic sperm injection. *Hum Reprod* 10:1554
44. Galli C, Vassiliev I, Lagutina I, Galli A, Lazzari G (2003) Bovine embryo development following ICSI: effect of activation, sperm capacitation and pre-treatment with dithiothreitol. *Theriogenology* 60:1467–1480
45. Catt SL, Catt JW, Gomez MC, Maxwell WM, Evans G (1996) Birth of a male lamb derived from an in vitro matured oocyte fertilised by intracytoplasmic injection of a single presumptive male sperm. *Vet Rec* 139:494–495
46. Kolbe T, Holtz W (2000) Birth of a piglet derived from an oocyte fertilized by intracytoplasmic sperm injection (ICSI). *Anim Reprod Sci* 64:97–101
47. Perry AC, Wakayama T, Kishikawa H, Kasai T, Okabe M, Toyoda Y, Yanagimachi R (1999) Mammalian transgenesis by intracytoplasmic sperm injection. *Science* 284:1180–1183
48. Gandolfi F, Moor RM (1987) Stimulation of early embryonic development in the sheep by co-culture with oviduct epithelial cells. *J Reprod Fertil* 81:23–28
49. Young LE, Sinclair KD, Wilmot I (1998) Large offspring syndrome in cattle and sheep. *Rev Reprod* 3:155–163
50. van Wagtenonk-de Leeuw AM, Mullaart E, de Roos AP, Merton JS, den Daas JH, Kemp B, de Ruigh L (2000) Effects of different reproduction techniques: AI MOET or IVP, on health and welfare of bovine offspring. *Theriogenology* 53:575–597
51. Tervit HR, Whittingham DG, Rowson LE (1972) Successful culture in vitro of sheep and cattle ova. *J Reprod Fertil* 30:493–497
52. Lazzari G, Wrenzycki C, Herrmann D, Duchi R, Kruip T, Niemann H, Galli C (2002) Cellular and molecular deviations in bovine in vitro-produced embryos are related to the large offspring syndrome. *Biol Reprod* 67:767–775

53. Gurdon JB (2006) From nuclear transfer to nuclear reprogramming: the reversal of cell differentiation. *Annu Rev Cell Dev Biol* 22:1–22
54. Willadsen SM (1986) Nuclear transplantation in sheep embryos. *Nature* 320:63–65
55. Galli C, Lagutina I, Crotti G, Colleoni S, Turini P, Ponderato N, Duchi R, Lazzari G (2003) Pregnancy: a cloned horse born to its dam twin. *Nature* 424:635
56. Lagutina I, Lazzari G, Duchi R, Colleoni S, Ponderato N, Turini P, Crotti G, Galli C (2005) Somatic cell nuclear transfer in horses: effect of oocyte morphology, embryo reconstruction method and donor cell type. *Reproduction* 130:559–567
57. Galli C, Duchi R, Moor RM, Lazzari G (1999) Mammalian leukocytes contain all the genetic information necessary for the development of a new individual. *Cloning* 1:161–170
58. Galli C, Duchi R, Crotti G, Turini P, Ponderato N, Colleoni S, Lagutina I, Lazzari G (2003) Bovine embryo technologies. *Theriogenology* 59:599–616
59. Kuroiwa Y, Kasinathan P, Matsushita H, Sathiyasalan J, Sullivan EJ, Kakitani M, Tomizuka K, Ishida I, Robl JM (2004) Sequential targeting of the genes encoding immunoglobulin- μ and prion protein in cattle. *Nat Genet* 36:775–780
60. Lombardo A, Genovese P, Beausejour CM, Colleoni S, Lee YL, Kim KA, Ando D, Urnov FD, Galli C, Gregory PD, Holmes MC, Naldini L (2007) Gene editing in human stem cells using zinc finger nucleases and integrase-defective lentiviral vector delivery. *Nat Biotechnol* 25:1298–1306
61. Wall RJ, Powell AM, Paape MJ, Kerr DE, Bannerman DD, Pursel VG, Wells KD, Talbot N, Hawk HW (2005) Genetically enhanced cows resist intramammary *Staphylococcus aureus* infection. *Nat Biotechnol* 23:445–451
62. Brophy B, Smolenski G, Wheeler T, Wells D, L'Huillier P, Laible G (2003) Cloned transgenic cattle produce milk with higher levels of beta-casein and kappa-casein. *Nat Biotechnol* 21:157–162
63. Wakayama T, Tabar V, Rodriguez I, Perry AC, Studer L, Mombaerts P (2001) Differentiation of embryonic stem cell lines generated from adult somatic cells by nuclear transfer. *Science* 292:740–743
64. Lazzari G, Colleoni S, Giannelli SG, Brunetti D, Colombo E, Lagutina I, Galli C, Broccoli V (2006) Direct derivation of neural rosettes from cloned bovine blastocysts: a model of early neurulation events and neural crest specification in vitro. *Stem Cells* 24:2514–2521
65. Takahashi K, Yamanaka S (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* 126:663–676
66. Evans MJ, Kaufman MH (1981) Establishment in culture of pluripotential cells from mouse embryos. *Nature* 292:154–156
67. Martin GR (1981) Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. *Proc Natl Acad Sci USA* 78:7634–7638
68. Keefer CL, Pant D, Blomberg L, Talbot NC (2007) Challenges and prospects for the establishment of embryonic stem cell lines of domesticated ungulates. *Anim Reprod Sci* 98:147–168
69. Notarianni E, Galli C, Laurie S, Moor RM, Evans MJ (1991) Derivation of pluripotent, embryonic cell lines from the pig and sheep. *J Reprod Fertil Suppl* 43:255–260
70. Iwasaki S, Campbell KH, Galli C, Akiyama K (2000) Production of live calves derived from embryonic stem-like cells aggregated with tetraploid embryos. *Biol Reprod* 62:470–475
71. Ying QL, Stavridis M, Griffiths D, Li M, Smith A (2003) Conversion of embryonic stem cells into neuroectodermal precursors in adherent monoculture. *Nat Biotechnol* 21:183–186
72. Ying QL, Wray J, Nichols J, Battle-Morera L, Doble B, Woodgett J, Cohen P, Smith A (2008) The ground state of embryonic stem cell self-renewal. *Nature* 453:519–523
73. Buehr M, Meek S, Blair K, Yang J, Ure J, Silva J, McLay R, Hall J, Ying QL, Smith A (2008) Capture of authentic embryonic stem cells from rat blastocysts. *Cell* 135:1287–1298
74. Tesar PJ, Chenoweth JG, Brook FA, Davies TJ, Evans EP, Mack DL, Gardner RL, McKay RD (2007) New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature* 448:196–199
75. Brons IG, Smithers LE, Trotter MW, Rugg-Gunn P, Sun B, de Chuva Sousa Lopes SM, Howlett SK, Clarkson A, Ahrlund-Richter L, Pedersen RA, Vallier L (2007) Derivation of pluripotent epiblast stem cells from mammalian embryos. *Nature* 448:191–195
76. Alberio R, Croxall N, Allegrucci C. Pig Epiblast Stem Cells Depend on Activin/Nodal Signaling for Pluripotency and Self Renewal. *Stem Cells Dev* 2010.
77. Brown BD, Gentner B, Cantore A, Colleoni S, Amendola M, Zingale A, Baccarini A, Lazzari G, Galli C, Naldini L (2007) Endogenous microRNA can be broadly exploited to regulate transgene expression according to tissue, lineage and differentiation state. *Nat Biotechnol* 25:1457–1467
78. Mansour SL, Thomas KR, Capecchi MR (1988) Disruption of the proto-oncogene int-2 in mouse embryo-derived stem cells: a general strategy for targeting mutations to non-selectable genes. *Nature* 336:348–352
79. Campbell KH, McWhir J, Ritchie WA, Wilmut I (1996) Sheep cloned by nuclear transfer from a cultured cell line. *Nature* 380:64–66
80. Galli C, Lazzari G (2008) The manipulation of gametes and embryos in farm animals. *Reprod Domest Anim* 43(Suppl 2):1–7
81. Brinster RL, Allen JM, Behringer RR, Gelinas RE, Palmiter RD (1988) Introns increase transcriptional efficiency in transgenic mice. *Proc Natl Acad Sci USA* 85:836–840
82. Palmiter RD, Sandgren EP, Avarbock MR, Allen DD, Brinster RL (1991) Heterologous introns can enhance expression of transgenes in mice. *Proc Natl Acad Sci USA* 88:478–482
83. Schedl A, Montoliu L, Kelsey G, Schutz G (1993) A yeast artificial chromosome covering the tyrosinase gene confers copy number-dependent expression in transgenic mice. *Nature* 362:258–261
84. Giraldo P, Montoliu L (2001) Size matters: use of YACs, BACs and PACs in transgenic animals. *Transgenic Res* 10:83–103

85. Van Keuren ML, Gavrilina GB, Filipiak WE, Zeidler MG, Saunders TL (2009) Generating transgenic mice from bacterial artificial chromosomes: transgenesis efficiency, integration and expression outcomes. *Transgenic Res* 18:769–785
86. Sauer B, Henderson N (1990) Targeted insertion of exogenous DNA into the eukaryotic genome by the Cre recombinase. *New Biol* 2:441–449
87. Allen GC, Spiker S, Thompson WF (2005) Transgene integration: use of matrix attachment regions. *Meth Molecular Biology Clifton N J* 286:313–326
88. Harraghy N, Gaussin A, Mermod N (2008) Sustained transgene expression using MAR elements. *Curr Gene Ther* 8:353–366
89. McKnight RA, Shamay A, Sankaran L, Wall RJ, Hennighausen L (1992) Matrix-attachment regions can impart position-independent regulation of a tissue-specific gene in transgenic mice. *Proc Natl Acad Sci USA* 89:6943–6947
90. McKnight RA, Spencer M, Wall RJ, Hennighausen L (1996) Severe position effects imposed on a 1 kb mouse whey acidic protein gene promoter are overcome by heterologous matrix attachment regions. *Mol Reprod Dev* 44:179–184
91. Polejaeva IA, Chen SH, Vaught TD, Page RL, Mullins J, Ball S, Dai Y, Boone J, Walker S, Ayares DL, Colman A, Campbell KH (2000) Cloned pigs produced by nuclear transfer from adult somatic cells. *Nature* 407:86–90
92. Geurts AM, Cost GJ, Freyvert Y, Zeitler B, Miller JC, Choi VM, Jenkins SS, Wood A, Cui X, Meng X, Vincent A, Lam S, Michalkiewicz M, Schilling R, Foeckler J, Kalloway S, Weiler H, Menoret S, Anegón I, Davis GD, Zhang L, Rebar EJ, Gregory PD, Urnov FD, Jacob HJ, Buelow R (2009) Knockout rats via embryo microinjection of zinc-finger nucleases. *Science* 325:433
93. Dickson D (1989) "Dangerous" liaisons in cell biology. *Science* 244:1539–1540
94. Chang K, Qian J, Jiang M, Liu YH, Wu MC, Chen CD, Lai CK, Lo HL, Hsiao CT, Brown L, Bolen J Jr, Huang HI, Ho PY, Shih PY, Yao CW, Lin WJ, Chen CH, Wu FY, Lin YJ, Xu J, Wang K (2002) Effective generation of transgenic pigs and mice by linker based sperm-mediated gene transfer. *BMC Biotechnol* 2:5
95. Shen W, Li L, Pan Q, Min L, Dong H, Deng J (2006) Efficient and simple production of transgenic mice and rabbits using the new DMSO-sperm mediated exogenous DNA transfer method. *Mol Reprod Dev* 73:589–594
96. Lavitrano M, Bacci ML, Forni M, Lazzereschi D, Di Stefano C, Fioretti D, Giancotti P, Marfe G, Pucci L, Renzi L, Wang H, Stoppacciaro A, Stassi G, Sargiacomo M, Sinibaldi P, Turchi V, Giovannoni R, Della Casa G, Seren E, Rossi G (2002) Efficient production by sperm-mediated gene transfer of human decay accelerating factor (hDAF) transgenic pigs for xenotransplantation. *Proc Natl Acad Sci USA* 99:14230–14235
97. Wu Z, Li Z, Yang J (2008) Transient transgene transmission to piglets by intrauterine insemination of spermatozoa incubated with DNA fragments. *Mol Reprod Dev* 75:26–32
98. Moreira PN, Perez-Crespo M, Ramirez MA, Pozueta J, Montoliu L, Gutierrez-Adan A (2007) Effect of transgene concentration, flanking matrix attachment regions, and RecA-coating on the efficiency of mouse transgenesis mediated by intracytoplasmic sperm injection. *Biol Reprod* 76:336–343
99. Kurome M, Ueda H, Tomii R, Naruse K, Nagashima H (2006) Production of transgenic-clone pigs by the combination of ICSI-mediated gene transfer with somatic cell nuclear transfer. *Transgenic Res* 15:229–240
100. Whitelaw CB, Radcliffe PA, Ritchie WA, Carlisle A, Ellard FM, Pena RN, Rowe J, Clark AJ, King TJ, Mitrophanous KA (2004) Efficient generation of transgenic pigs using equine infectious anaemia virus (EIAV) derived vector. *FEBS Lett* 571:233–236
101. Hacein-Bey-Abina S, Garrigue A, Wang GP, Soulier J, Lim A, Morillon E, Clappier E, Caccavelli L, Delabesse E, Beldjord K, Asnafi V, MacIntyre E, Dal Cortivo L, Radford I, Brousse N, Sigaux F, Moshous D, Hauer J, Borkhardt A, Belohradsky BH, Wintergerst U, Velez MC, Leiva L, Sorensen R, Wulffraat N, Blanche S, Bushman FD, Fischer A, Cavazzana-Calvo M (2008) Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of SCID-X1. *J Clin Invest* 118:3132–3142
102. Howe SJ, Mansour MR, Schwarzwald K, Bartholomae C, Hubank M, Kempinski H, Brugman MH, Pike-Overzet K, Chatters SJ, de Ridder D, Gilmour KC, Adams S, Thornhill SI, Parsley KL, Staal FJ, Gale RE, Linch DC, Bayford J, Brown L, Quayle M, Kinnon C, Ancliff P, Webb DK, Schmidt M, von Kalle C, Gaspar HB, Thrasher AJ (2008) Insertional mutagenesis combined with acquired somatic mutations causes leukemia following gene therapy of SCID-X1 patients. *J Clin Invest* 118:3143–3150
103. Romano G (2009) An update on gene therapy programs. *Drug News Perspect* 22:435–440
104. de Sio FR, Santoni, Gritti A, Cascio P, Neri M, Sampaoli M, Galli C, Luban J, Naldini L (2008) Lentiviral vector gene transfer is limited by the proteasome at postentry steps in various types of stem cells. *Stem Cells* 26:2142–2152
105. Chan AW, Chong KY, Martinovich C, Simerly C, Schatten G (2001) Transgenic monkeys produced by retroviral gene transfer into mature oocytes. *Science* 291:309–312
106. Hofmann A, Kessler B, Ewerling S, Kabermann A, Brem G, Wolf E, Pfeifer A (2006) Epigenetic regulation of lentiviral transgene vectors in a large animal model. *Mol Ther* 13:59–66
107. Henckaerts E, Dutheil N, Zeltner N, Kattman S, Kohlbrenner E, Ward P, Clement N, Rebollo P, Kennedy M, Keller GM, Linden RM (2009) Site-specific integration of adeno-associated virus involves partial duplication of the target locus. *Proc Natl Acad Sci USA* 106:7571–7576
108. Miller DG, Wang PR, Petek LM, Hirata RK, Sands MS, Russell DW (2006) Gene targeting in vivo by adeno-associated virus vectors. *Nat Biotechnol* 24:1022–1026

109. Russell DW, Hirata RK (1998) Human gene targeting by viral vectors. *Nat Genet* 18:325–330
110. Rogers CS, Stoltz DA, Meyerholz DK, Ostedgaard LS, Rokhlina T, Taft PJ, Rogan MP, Pezzulo AA, Karp PH, Itani OA, Kabel AC, Wohlford-Lenane CL, Davis GJ, Hanfland RA, Smith TL, Samuel M, Wax D, Murphy CN, Rieke A, Whitworth K, Uc A, Starner TD, Brogden KA, Shilyansky J, McCray PB Jr, Zabner J, Prather RS, Welsh MJ (2008) Disruption of the CFTR gene produces a model of cystic fibrosis in newborn pigs. *Science* 321:1837–1841
111. Rogers CS, Hao Y, Rokhlina T, Samuel M, Stoltz DA, Li Y, Petroff E, Vermeer DW, Kabel AC, Yan Z, Spate L, Wax D, Murphy CN, Rieke A, Whitworth K, Linville ML, Korte SW, Engelhardt JF, Welsh MJ, Prather RS (2008) Production of CFTR-null and CFTR-DeltaF508 heterozygous pigs by adeno-associated virus-mediated gene targeting and somatic cell nuclear transfer. *J Clin Invest* 118:1571–1577
112. Brunetti D, Perota A, Lagutina I, Colleoni S, Duchi R, Calabrese F, Seveso M, Cozzi E, Lazzari G, Lucchini F, Galli C (2008) Transgene expression of green fluorescent protein and germ line transmission in cloned pigs derived from in vitro transfected adult fibroblasts. *Cloning Stem Cells* 10(4):409–420
113. Lagutina I, Lazzari G, Duchi R, Turini P, Tessaro I, Brunetti D, Colleoni S, Crotti G, Galli C (2007) Comparative aspects of somatic cell nuclear transfer with conventional and zona-free method in cattle, horse, pig and sheep. *Theriogenology* 67:90–98
114. Lagutina I, Lazzari G, Galli C (2006) Birth of cloned pigs from zona-free nuclear transfer blastocysts developed in vitro before transfer. *Cloning Stem Cells* 8:283–293
115. Kragh PM, Du Y, Corydon TJ, Purup S, Bolund L, Vajta G (2005) Efficient in vitro production of porcine blastocysts by hand-made cloning with a combined electrical and chemical activation. *Theriogenology* 64:1536–1545
116. Vajta G, Kragh PM, Mtango NR, Callesen H (2005) Hand-made cloning approach: potentials and limitations. *Reprod Fertil Dev* 17:97–112
117. Esteban MA, Xu J, Yang J, Peng M, Qin D, Li W, Jiang Z, Chen J, Deng K, Zhong M, Cai J, Lai L, Pei D (2009) Generation of induced pluripotent stem cell lines from Tibetan miniature pig. *J Biol Chem* 284:17634–17640
118. Ezashi T, Telugu BP, Alexenko AP, Sachdev S, Sinha S, Roberts RM (2009) Derivation of induced pluripotent stem cells from pig somatic cells. *Proc Natl Acad Sci USA* 106:10993–10998
119. Wu Z, Chen J, Ren J, Bao L, Liao J, Cui C, Rao L, Li H, Gu Y, Dai H, Zhu H, Teng X, Cheng L, Xiao L (2009) Generation of pig induced pluripotent stem cells with a drug-inducible system. *J Mol Cell Biol* 1:46–54
120. Takahagi Y, Fujimura T, Miyagawa S, Nagashima H, Shigehisa T, Shirakura R, Murakami H (2005) Production of alpha 1, 3-galactosyltransferase gene knockout pigs expressing both human decay-accelerating factor and N-acetylglucosaminyltransferase III. *Mol Reprod Dev* 71:331–338
121. Fujimura T, Takahagi Y, Shigehisa T, Nagashima H, Miyagawa S, Shirakura R, Murakami H (2008) Production of alpha 1, 3-galactosyltransferase gene-deficient pigs by somatic cell nuclear transfer: a novel selection method for gal alpha 1, 3-Gal antigen-deficient cells. *Mol Reprod Dev* 75:1372–1378
122. Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, Yamanaka S (2007) Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* 131:861–872
123. Yu J, Vodyanik MA, Smuga-Otto K, Antosiewicz-Bourget J, Frane JL, Tian S, Nie J, Jonsdottir GA, Ruotti V, Stewart R, Slukvin II, Thomson JA (2007) Induced pluripotent stem cell lines derived from human somatic cells. *Science* 318:1917–1920
124. Huangfu D, Osafune K, Maehr R, Guo W, Eijkelenboom A, Chen S, Muhlestein W, Melton DA (2008) Induction of pluripotent stem cells from primary human fibroblasts with only Oct4 and Sox2. *Nat Biotechnol* 26:1269–1275
125. Feng B, Ng JH, Heng JC, Ng HH (2009) Molecules that promote or enhance reprogramming of somatic cells to induced pluripotent stem cells. *Cell Stem Cell* 4:301–312
126. Ralston A, Rossant J (2010) The genetics of induced pluripotency. *Reproduction* 139:35–44
127. Bartolome C, Bello X, Maside X (2009) Widespread evidence for horizontal transfer of transposable elements across *Drosophila* genomes. *Genome Biol* 10:R22
128. Clark KJ, Carlson DF, Foster LK, Kong BW, Foster DN, Fahrenkrug SC (2007) Enzymatic engineering of the porcine genome with transposons and recombinases. *BMC Biotechnol* 7:42
129. Araki K, Araki M, Yamamura K (2002) Site-directed integration of the cre gene mediated by Cre recombinase using a combination of mutant lox sites. *Nucleic Acids Res* 30:e103
130. Irion S, Luche H, Gadue P, Fehling HJ, Kennedy M, Keller G (2007) Identification and targeting of the ROSA26 locus in human embryonic stem cells. *Nat Biotechnol* 25:1477–1482
131. Nyabi O, Naessens M, Haigh K, Gembarska A, Goossens S, Maetens M, De Clercq S, Drogat B, Haenebalcke L, Bartunkova S, De Vos I, De Craene B, Karimi M, Berx G, Nagy A, Hilson P, Marine JC, Haigh JJ (2009) Efficient mouse transgenesis using Gateway-compatible ROSA26 locus targeting\ vectors and F1 hybrid ES cells. *Nucleic Acids Res* 37:e55
132. Porteus MH (2006) Mammalian gene targeting with designed zinc finger nucleases. *Mol Ther* 13:438–446
133. d'Apice AJ, Cowan PJ (2008) Gene-modified pigs. *Xenotransplantation* 15:87–90
134. Fang J, Qian JJ, Yi S, Harding TC, Tu GH, VanRoey M, Jooss K (2005) Stable antibody expression at therapeutic levels using the 2A peptide. *Nat Biotechnol* 23:584–590

135. Bartosch B, Stefanidis D, Myers R, Weiss R, Patience C, Takeuchi Y (2004) Evidence and consequence of porcine endogenous retrovirus recombination. *J Virol* 78:13880–13890
136. Moalic Y, Felix H, Takeuchi Y, Jestin A, Blanchard Y (2009) Genome areas with high gene density and CpG island neighborhood strongly attract porcine endogenous retrovirus for integration and favor the formation of hot spots. *J Virol* 83:1920–1929
137. Takeuchi Y, Patience C, Magre S, Weiss RA, Banerjee PT, Le Tissier P, Stoye JP (1998) Host range and interference studies of three classes of pig endogenous retrovirus. *J Virol* 72:9986–9991
138. Dieckhoff B, Petersen B, Kues WA, Kurth R, Niemann H, Denner J (2008) Knockdown of porcine endogenous retrovirus (PERV) expression by PERV-specific shRNA in transgenic pigs. *Xenotransplantation* 15:36–45
139. Kragh PM, Nielsen AL, Li J, Du Y, Lin L, Schmidt M, Bogh IB, Holm IE, Jakobsen JE, Johansen MG, Purup S, Bolund L, Vajta G, Jorgensen AL (2009) Hemizygous minipigs produced by random gene insertion and handmade cloning express the Alzheimer's disease-causing dominant mutation APPsw. *Transgenic Res* 18:545–558
140. Petersen B, Carnwath JW, Niemann H (2009) The perspectives for porcine-to-human xenografts. *Comp Immunol Microbiol Infect Dis* 32:91–105
141. Galli C, Perota A, Brunetti D, Lagutina I, Lazzari G, Lucchini F (2010) Genetic engineering including superseding micro-injection: new ways to make GM pigs. *Xenotransplantation* 17(6):397–410
142. Lai L, Kang JX, Li R, Wang J, Witt WT, Yong HY, Hao Y, Wax DM, Murphy CN, Rieke A, Samuel M, Linville ML, Korte SW, Evans RW, Starzl TE, Prather RS, Dai Y (2006) Generation of cloned transgenic pigs rich in omega-3 fatty acids. *Nat Biotechnol* 24:435–436

Books

- Gordon IR (2003) Laboratory production of cattle embryos, vol Series No. 27, 2nd edn, Biotechnology in agriculture. CAB International, Wallingford
- Gordon I (2004) Reproductive technologies in farm animals. CAB International, Wallingford. ISBN 0 85199 862 3
- Hafez B, Hafez ESE (2000) Reproduction in farm animals, 7th ed. Lippincott Williams & Wilkins. ISBN 0-683-30577-8
- Monk M (1987) Mammalian development, a practical approach. In: Rickwood D, Hames BD (eds) The practical approach series. IRL Press Ltd, Oxford/Washington, DC
- Nagy A (2003) Manipulating the mouse embryo : a laboratory manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor. ISBN: 0879695749 0879695919
- Robertson E (1987) Teratocarcinoma and embryonic stem cells, a practical approach. In: Rickwood D, Hames BD (eds) The practical approach series. IRL Press Ltd, Oxford/Washington, DC

Transgenic Livestock, Decreasing Environmental Impact of

SERGUEI P. GOLOVAN¹, CECIL W. FORSBERG²

¹Department of Animal and Food Science, University of Delaware, Newark, DE, USA

²Department of Molecular and Cellular Biology, University of Guelph, Guelph, ON, Canada

Article Outline

Glossary
Definition of the Subject
Introduction
Future Directions
Bibliography

Glossary

Arabinoxylan A polymeric backbone of $\beta(1,4)$ linked xylose residues with attached C(O)-2,3-linked arabinose residues that is found mainly in cereal grains. It becomes highly viscous when dissolved in the gastrointestinal tract and is not digested by monogastric animals.

Eutrophication A process where water bodies receive excess nutrients especially as phosphorus or nitrogen that stimulate excessive plant and algal growth resulting in reduced water quality.

β -Glucan A polymer found primarily in cereal grains consisting of $\beta(1,3;1,4)$ linked glucose residues that becomes viscous on solubilization in the gastrointestinal tract and is not digested by monogastric animals.

Glycanase Any enzyme that catalyzes the hydrolysis of a glycan, which includes glucanases and xylanases of all types.

Phytase Any type of phosphatase enzyme that catalyzes the hydrolysis of phytic acid releasing phosphate molecules.

Phytic Acid Inositol hexakisphosphate (or phytate when in the salt form) is the principle storage form of phosphorus in many plants. It accounts for 50–80% of the total phosphorus present and is poorly digested by monogastric animals.

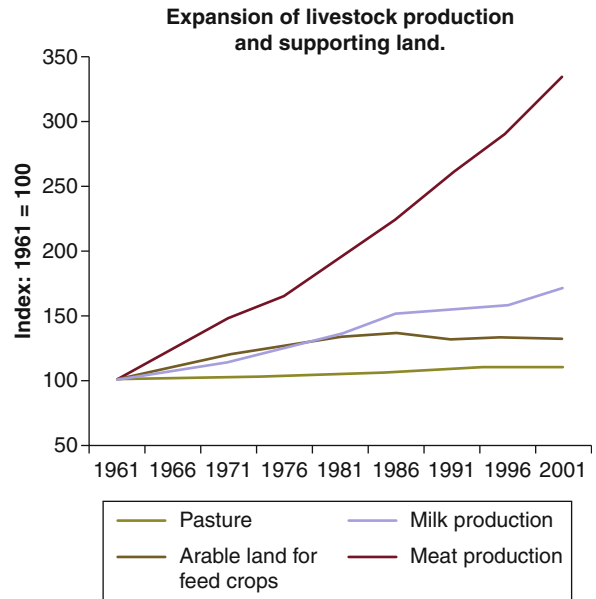
Definition of the Subject

As the world population increases, there is an increasing demand for food including both plant and animal products. The demand for meat and milk are increasing at a faster rate than for plant products because of increased wealth in developing countries. Intensification of livestock production to satisfy the demand is exacerbating the deleterious impact of intensive animal agriculture on the environment and new approaches are needed to reduce the impact. One approach is through the development of transgenic animals that have a smaller environmental footprint. At present there are no transgenic livestock in production. Transgenic livestock that have reduced environmental impact have been developed, but most proposed strategies are at the early stages of development. Physiological modifications that would reduce the impact of livestock include changes to improve feed utilization, increase growth and increase disease resistance, reduce manure output, and decrease greenhouse gas production. Obviously an essential factor is that the changes in physiology of the transgenic animal must have no deleterious effect on health, welfare, performance, and the environment.

Introduction

Environmental Impact of Livestock

Demand projections point to increases of global meat consumption of 68% and of global milk consumption of 57% of the 2000 base period by 2030 [26, 83]. While agriculture could be considered a “green” industry which uses solar energy to produce food and fiber for human use, it has also become a serious threat to global ecological systems. Currently forest losses in tropical countries is largely due to conversion to agricultural land to increase global grazing and cropland dedicated to the production of feed that already amounts to approximately 70% of all agricultural land [84]. The trends are illustrated by Fig. 1. Livestock production is a major cause of greenhouse gas production, particularly methane production, and as well is a major source of pollution due to the extensive use of pesticides, antibiotics, and nutrients (nitrogen, phosphorus, and minerals) that result in increased eutrophication with deterioration in water quality in



Steinfeld H, Gerber P PNAS 2010;107:18237–18238

Transgenic Livestock, Decreasing Environmental Impact of. Figure 1

Global expansion of livestock production and supporting land

many countries [87]. The major challenge for animal agriculture in this century is to sustain and increase meat and milk production without further degrading the environment.

Production of meat from monogastric animals, mainly swine and poultry, has increased 103% over the period from 1987 to 2007 compared to ruminant meat production that increased by 28% [27]. The shift to monogastric species and continued productivity gains have the potential to reduce environmental impact per unit of animal production in the future, but with overall production increasing, serious pollution concerns remain. The issues and options in addressing the environmental consequences of a growing livestock sector are clearly enunciated in a paper by Gerber et al. [30].

Selective breeding of animals has greatly improved commercial performance of livestock and is especially useful for improvement of *multi-gene traits*. Yet, selective breeding is a slow process, depending on the rate of animal reproduction and can only operate on alleles and genes present in animal population with little control over what to express, where, when, and how

much. Continuous selection for a particular trait may result in decreased genetic variability in the locus and decrease ability for further selection. For example, only a few proven bulls produce the majority of offspring in the world Holstein populations and similar reduced breed diversity in US swine and poultry is a cause for concern [56]. It is also difficult to introduce beneficial alleles from rare breeds into existing commercial population because it leads to decreased performance for economically important traits. In contrast, animal genetic engineering allows precise introduction of new alleles from rare breeds, novel genes from evolutionary distant species, or even completely artificial genes with control of the time, place, and amount of gene expression in an animal. Furthermore, a transgenic animal is a self-replicating high-tech animal – it might be expensive to produce but it is cheap to propagate.

Reducing the impact on the environment through animal transgenesis can take many forms (Table 1), for example, improving the overall feed digestion by the

animal or reducing the excretion of a specific nutrient such as phosphorus. More subtle means of improving the efficiency of animal production could include enhanced disease resistance or even improving the nutritional quality of the meat since it would increase the product value.

Swine

Growth and Improved Digestion The first transgenic food animal developed that exhibited a trait consistent with a reduced environmental footprint was the transgenic line of pigs developed by Pursel and colleagues [64]. These pigs contained an increased level of bovine growth hormone that gave rise to an 18% increase in feed efficiency and an 11–14% increase in growth rate (Table 1). These improvements obviously decreased manure production per unit of body weight gain. Unfortunately, the pigs exhibited lameness, stress susceptibility, gastric ulcers, and other

Transgenic Livestock, Decreasing Environmental Impact of. Table 1 Genetic modifications in livestock to improve overall performance and reduce environmental impact

Species	Trait	Protein/RNA; Transgene	Reference
Pig	Phosphorus metabolism	Phytase; PSP-APPA	[33]
Pig	Increased growth	Growth hormone; hMTpGH	[57]
Pig	Increased growth	Insulin-like growth factor; (mMT-HIGF-1)	[65]
Pig	Increased lactose in milk	α -lactalbumin; bovine α -lactalbumin	[90]
Pig	Influenza resistance	Mx protein; mMx1-Mx	[53]
Pig, Cattle	Disease resistance	IgA; mouse α and χ chains	[46]
Cattle	Avoids spontaneous prion occurrence	Knockout vectors pBPrP(H)KOneo; pBPrP(H)KOpuro	[73]
Cattle	<i>Staphylococcus aureus</i> resistance	Lysostaphin; β -lactoglobulin promoter linked to lysostaphin, neomycin resistance, and GFP	[88]
Cattle	Mastitis resistance	Human lactoferrin in milk	[79]
Goat	Reduced mastitis and healthier kids	Human lysozyme; A _{s1} -casein promoter-hLZcDNA	[38]
Sheep	Visna virus envelope	Visna LTR-env	[14]
Sheep	Ovine prion locus	Homologous recombination	[21]
Sheep	Increased growth	Ovine growth hormone; metallothionein promoter	[1]
Chicken	Influenza virus resistance	Chicken U6 promoter; cRNA binding site influenza virus polymerase	[49]

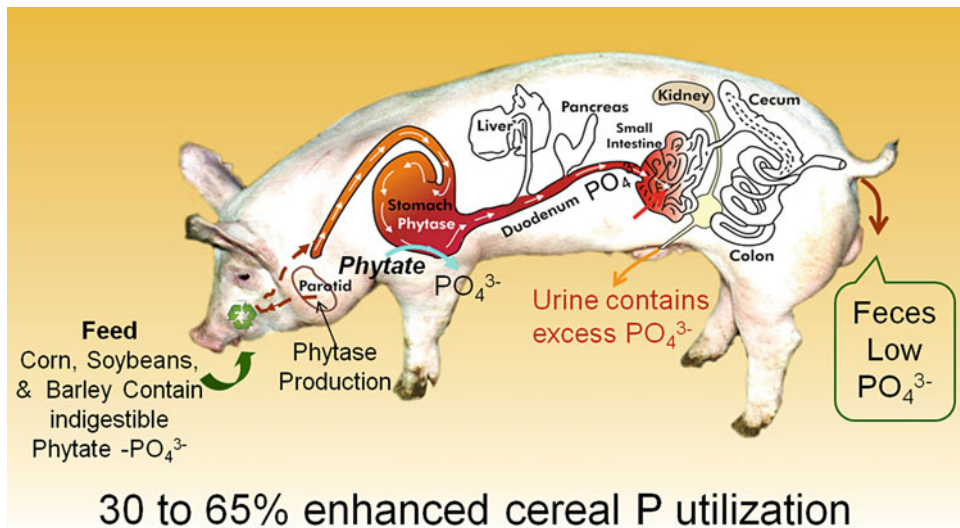
health problems that negated the unique improvements observed, but the study documented that the problem was due to overproduced growth hormone. In a separate study by Nottle et al. [57], the level of growth hormone was controlled by using an inducible promoter, a modified human metallothionein promoter fused to the cDNA sequence for the porcine growth hormone. They documented improved growth of founder animals as compared to non-transgenic littermates, but the study was not carried beyond littermates. Later, Pursel and colleagues [63] developed transgenic pigs that expressed insulin-like growth factor I (IGF-I) to determine whether directing the expression of IGF-I specifically to striated muscle would enhance lean muscle growth in pigs. However, they observed only a decreased daily rate of fat accretion in the transgenic pigs as compared to the rate in the unmodified pigs. This may be explained by the observation that transgenic modification of a trait that has already undergone intensive selection through traditional breeding and is a metabolic process with multiple effects can have unexpected and in some cases deleterious outcomes [22].

A different approach to enhancing growth rate of pigs was achieved by the introduction of the bovine α -lactalbumin gene into pigs. Sows containing the α -lactalbumin gene produce an increased concentration of lactose in milk during the early lactation that increases the growth rates of piglets [55, 90]. This novel transgenic modification improves efficiency of production by either allowing earlier weaning or more robust piglets at weaning.

Phosphorus is a key nutrient for all living things. At the same time, phosphorus mining damages the environment, and phosphorus-induced eutrophication leads to harmful algal blooms, decreased biodiversity, and changes in species composition. In addition, the global supply of mined phosphorus is running out and some estimate that within 30–40 years there will not be sufficient phosphorus to meet agricultural demand [16, 76]. Manure is a valuable organic fertilizer, but the high phosphorus (P) concentration in relation to the nitrogen (N) content is problematic. A desirable N/P ratio for plant growth is 6.0 for an increment in yield, [42]. This ratio is derived by using the N and P concentrations after losses due to volatilization, which is primarily an N loss. In contrast, the N/P values for

manure from pigs, dairy cattle, and poultry are approximately 0.96, 2.6, and 1.7, respectively. Therefore manures from pigs, dairy cattle, and poultry are 6.3, 2.3, and 3.4-fold enriched in phosphorus in relation to nitrogen utilized by cereal grain crops. This documents the doubly serious impact of excess P relative to N in the manure for land spreading. To decrease phosphorus-induced eutrophication caused by the high phosphorus content of pig manure and to reduce the need for supplemental phosphorus in the diet, Golovan and colleagues [33] developed transgenic pigs expressing an *Escherichia coli* phytase driven by the mouse parotid secretory protein promoter that gave rise to phytase production in the salivary glands (Fig. 2). Phytase, secreted by the parotid gland in the saliva, mixes with the incoming food particles during chewing and hydrolyzes phytate in the acidic environment of the stomach-releasing phosphate that is readily absorbed in the small intestine. Since the dietary phosphorus requirement of the transgenic pigs can be satisfied by the cereal grain diet without inclusion of either supplemental P or supplemental microbial phytase, there is an overall decrease in P excretion in the feces and urine that will enter the environment [33]. A line of these phytase pigs currently in the eighth generation exhibited salivary phytase activities at a levels similar to that of the founding transgenic pig, and excrete 30–65% less phosphorus in the manure depending upon the stage of growth and diet consumed. Trials have documented that the phytase pigs have similar reproductive characteristics, health, and growth rates to that of conventional pigs (Forsberg, unpublished data). The pig parotid secretory protein promoter was also tested [95], but the salivary phytase production in mice was not as efficient as that described by Golovan et al. [32] for the mouse PSP promoter. It would be very instructive to assess the efficacy of the pig PSP promoter driving phytase synthesis in the pig.

Cereal grains and plant protein supplements have an imbalance of amino acids with an excess of nonessential amino acids and lesser amounts of the essential amino acids lysine methionine and threonine [58]. This imbalance results in high excretion of nitrogen from nonessential amino acids, and the swine ammonia emission metric (kg livestock^{-1}) is higher than either beef or dairy cattle [10]. Ammonia emissions are substantially reduced by inclusion in the diet of



Transgenic Livestock, Decreasing Environmental Impact of. Figure 2
High phytate feed consumption by the Enviropig™

essential amino acids lysine, methionine, and threonine. Endowing pigs with the selected ability to synthesize essential amino acids in the appropriate proportions would decrease the need for protein supplements and would also decrease the excretion of ammonia arising from degradation of nonessential amino acids. To test this thesis, Rees and Hay [71] genetically modified mouse 3T3 cells by the introduction of a chimeric gene containing the coding regions of the *E. coli* gene for aspartokinase I/homoserine dehydrogenase I and the *Corynebacterium glutamicum* gene for aspartic semialdehyde dehydrogenase subcloned into a Simian virus 40 based mammalian expression vector. These cells produced homoserine. By transfecting these cells with the plasmid pSVthrB/c containing genes coding for homoserine kinase and threonine synthase, the modified cell line expressed the complete pathway for the synthesis of threonine from aspartic acid, and the cell line no longer had a growth requirement for threonine. Therefore, by using the appropriate promoters with these genes it should be possible to produce genetically modified pigs/animals with the innate endogenous capacity to synthesize the essential amino acid threonine. The introduction of genes for the endogenous synthesis of methionine and lysine would be highly beneficial; unfortunately, the development of transgenes appropriately regulated to make each

biosynthetic pathway function in tissues will be a daunting task because methionine synthesis requires four additional genes coding for enzymes that convert of homoserine to methionine. Lysine synthesis would require as many as nine genes coding for enzymes in the pathway beginning with aspartic acid (<http://www.genome.jp/kegg/pathway/map/map00300>). Alternatively, it might be possible to improve utilization of essential amino acids present in feed by increasing ileal digestibility (phytase, cellulase), absorption from gut (amino acid transporters), or even by increasing proportion of gut microorganisms responsible for synthesis of the essential amino acids by developing selective attachment receptors or by suppression of competitors using antimicrobial proteins.

Cereal grains and plant protein supplements contain indigestible structural carbohydrate components including β -glucan and arabinoxylan that are not digested by nonruminant species and are mainly excreted in the manure. It has been shown that supplementation of the diet with exogenous β -glucanase and xylanase improves growth of the piglets [23]. However, supplementation of grower diets with glycanase enzymes has given mixed results. In a study by Nyachoti et al. [59] no beneficial effect was observed by the inclusion of glycanase while Ji et al. [40] observed a distinct beneficial effect on digestion.

The beneficial effect of supplemental glucanase was the basis for testing whether this class of enzymes can be expressed in animals to enhance digestion. To test this hypothesis a transgene composed of the mouse pancreas-specific amylase 2.2 promoter and the associated signal sequence was linked to the *Bacillus subtilis* endoglucanase with a 3' polyadenylation sequence. The transgene was introduced into mice by pronuclear microinjection. Offspring containing the transgene expressed the endoglucanase gene in the pancreas and secreted the truncated, but active glucanase into the small intestine [96]. Ohnishi et al. [60] showed that overexpression of the Rab3D upregulated amylase secretion from the pancreatic acini of transgenic mice; therefore it is possible this technique could be used to further enhance glycanase synthesis and secretion from the pancreas into the small intestine. Together with a report by Hall et al. [35] describing the transgenic expression of a *Clostridium thermocellum* thermostable glucanase in the pancreas of the mouse using an elastase promoter/enhancer, these studies show the feasibility of expressing hydrolase genes in the gastrointestinal tracts of livestock.

Health Status Health is a major issue in livestock production documented by the deleterious economic impacts [72] of current and emerging swine zoonoses [82]. Strategies to improve resistance of swine to viral diseases such as foot-and-mouth disease, influenza, and porcine reproductive and respiratory syndrome which depend on host cellular machinery include antisense, decoys, ribozyme, and RNA interference [47, 92]. Recent work on the control of the African swine fever virus through the use of RNA interference may soon show success [41]. To provide resistance against a broad range of bacteria and some viruses, Cheung et al. [11] introduced into mice the porcine protegrin-1 using the cytomegalovirus promoter. Protegrin is an antimicrobial peptide that targets both gram-negative and gram-positive bacteria as well as enveloped viruses. They showed increased resistance to the swine pathogen *Actinobacillus suis*. This protein normally is expressed in neutrophils, and the intent in this study was to assess the effect of more general expression that would affect the bacteria at an earlier stage in the infection. Introduction of this construct into pigs and subsequent testing will be very

informative since subclinical infections in pigs are a major cause of poorer feed efficiency in many environments.

Ruminants

Enhanced Digestion No transgenic ruminants have been reported that have a reduced impact on the environment through enhanced growth characteristics or reduced nutrient excretion. However, it was shown that feeding phytase and cellulases to lactating dairy cows resulted in reduced fecal excretion of dry matter, neutral detergent fiber, and acid detergent fiber, and reduced nitrogen and phosphorus in feces [43]. The beneficial effect of supplemental enzymes was found despite the presence of 101 distinct ruminal microbial phytases [37] and 27,755 putative carbohydrate-active genes, many of which presumably code for ruminal plant cell wall degrading enzymes [36]. Because added phytase enhances phosphorus utilization, transgenic expression of a phytase in the salivary glands of the cow using a promoter such as the indigenous salivary Bsp30a protein promoter [66, 91] would seem to be a possible strategy since Bsp30a is produced at a high concentration in saliva. No information is available on the Bsp30a promoter at this time; therefore background work would be necessary before testing this hypothesis.

Whether there would be interest in exploring the salivary production of glycanases in ruminants is questionable since variable results have been obtained by supplementation of diets with cellulases [5, 6, 62].

Methane from livestock accounts for approximately 37% of the methane produced by human-related activities, and the single largest source being enteric fermentation, mainly in ruminant livestock [30]. To reduce methane production in cattle, one option proposed was to generate transgenic cattle that produce salivary antibodies against rumen methanogens which subsequently bind to and inhibit their action in the rumen [44]. The same authors also identified the prophage ϕ -mru that is able to lyse methanogen cells. Secretion of anti-methanogenic proteins in the saliva would seem feasible from a physiological perspective as there are several strong salivary promoters [91]. However, before such experiments are undertaken it obviously would

be prudent to conduct feeding trials and testing for effects on the methanogenic population and to monitor the effect on the animal.

Health Disease is an ongoing drag on the efficiency in the production of ruminant animals [72]. Mastitis has a serious impact on milk production by ruminants [4]. To eliminate the deleterious effect of mastitis in dairy cows, Wall et al. [88] expressed lysostaphin in the milk. This antimicrobial protein enhanced the resistance to *Staphylococcus aureus* and provisionally could help maintain high milk production as a consequence of reduced mastitis if the genetic modification were accepted for commercial production. In a similar vein Maga et al. [51] demonstrated that human lysozyme expressed in the mammary gland of transgenic dairy goats inhibited the growth of bacteria that cause mastitis and this occurs without an effect on health or performance [38]. In a different approach Simojoki et al. [80] genetically modified dairy cows to secrete human lactoferrin in the milk. Again, it enhanced resistance to mastitis infections. Other avenues to improved performance could be through prion knock-outs [73] and developing resistance to viruses (infectious bovine rhinotracheitis, bovine respiratory syncytial viruses, parainfluenza, bovine viral diarrhea, rabies, and foot-and-mouth disease) through the application of an RNA interference-based approach [47].

Poultry

Enhanced Digestion Inclusion of phytase, glucanase, and xylanase in the poultry diet enhances feed digestibility. Phytase was shown to enhance availability of phosphorus and other minerals, and in addition, to enhance amino acid digestibility [18]. The chicken is reported to have a weak endogenous magnesium-stimulated phytase activity associated with the intestinal brush boarder [50], but apparently contributes little to overall phytate digestion. Initial work has been done to develop a secretory competent form of the enzyme with higher activity for expression in poultry [12], although no further work has been reported. Use of the salivary glands of poultry as a site for transgenic phytase production is a possibility since these glands have been identified [61, 75]. However, based on the low amylase activity of the salivary glands reported for

both chickens and turkeys [39], it is questionable whether a suitable promoter is available and a sufficient capacity for synthesis is possible to provide the quantity of enzyme necessary. Another possible site for secretion of hydrolytic enzymes is in the proventriculus using the chitinase promoter [85]; however, this would initially require the cloning and characterization of the promoter. Introduction of any transgene probably would involve use of either a combinatorial cis-regulatory element [77] or a lentiviral vector [49].

Arabinoxylanases alleviate viscosity-induced diffusion constraints associated with diets containing wheat, rye, barley, and triticale. Glucanases have a similar effect digesting β -glucans present and small amounts of amorphous cellulose. These glycanases were shown to enhance the amino acid digestibility [18]. The positive action of these feed enzymes opens the possibility for genetic modification.

Health Subclinical disease in poultry dramatically increases the environmental impact of the birds. For example, subclinical necrotic enteritis in broiler chickens results in a 12% reduction in body weight and an 11% decrease in feed utilization efficiency [81]. Other subclinical diseases undoubtedly have a similar impact; therefore, improving health substantively reduced the environmental footprint of poultry. The application of interference-based gene silencing may be an effective strategy for control of Marek's disease, infectious bursal disease, avian leucosis, Rous sarcoma virus, and avian influenza [47]. The report by Lyall et al. [49] on the production of transgenic chickens resistant to avian influenza viruses is an excellent example of the use of RNA interference to reduce mortalities of producing birds.

Issues with Expression of Novel Hydrolases in the Gastrointestinal Tract

Novel gene products that either have been expressed in the gastrointestinal tract of mammals, or that may be considered for expression are summarized in Table 2. Many phytases have been characterized and considered as potential feed enzymes [67], although only three phytases are illustrated as examples, the *E. coli* phytase expressed in the EnviropigTM, the Avian phytase and the rat phytase, the only phytases so far identified in

Transgenic Livestock, Decreasing Environmental Impact of. Table 2 Enzymes with some criteria suitable for gastrointestinal expression

Source	Catalytic properties	Mol. Wt.	pH Opt and Min/Max	Temp (°C) Max Act.	Sp.Act. U/mg ⁻¹ min ⁻¹	Protease resistance
<i>E. coli</i> [31]	Phytase/acid phosphatase	44,708 Da	Opt 4.5/2.5-3 Min 2.0 Max 10	60	2,326/712	Pepsin resistant and trypsin sensitive
Avian histidine phosphatase [74], [12]	Acid phosphatase/phytase	53,000 and 55,000 Da	4.5-7.5	Assay 40	0.5	?
Rat [94]	Phytase and alkaline phosphatase	70,000 and 90,000 Da	Phytase – 7.5 Phosphatase – 10.4	Assay 37	5.6	?
<i>Bacillus</i> [70]	Bifunctional phytase/endoglucanase	73,000 Da	Phytase – 7 Endoglucanase – 6 Range 4.0 – 8.0	55	Phytase – 8 Endoglucanase – 10	?
<i>F. succinogenes</i> [89]	1,3-1,4-β-glucanase	27,957 Da	Opt 6-8 Min 4.0	50	10,800	?
<i>Bisporus</i> sp. [48]	1,3-1,4-β-glucanase	48,000 Da after deglycosylation	Opt 5.0 Range 1-8	60	4,040	Pepsin and trypsin resistant
<i>Alicyclobacillus</i> sp. [2]	Endo-β-1,4-glucanase	48,000 Da	Opt 2.6 Range 1.8-7.6	65	5,032	Pepsin and trypsin resistant
<i>Alicyclobacillus</i> sp. [3]	Endo-β-1,4-glucanase	60,000 Da	3.4 Range 1.2-8.2	60	?	Pepsin and trypsin resistant
<i>Aspergillus niger</i> [45]	Endo-β-1,4-xylanase	23,000 Da	Opt 5.5	50	3,881	?
<i>N. patriciarum</i> [93]	Endo-β-1,4-xylanase	26,000 Da	Opt 6.0	60	5,000	?
<i>Phialophora</i> sp. G5 [97]	Endo-β-1,4-xylanase	55,000 Da	Opt 4.0 Range	70	351	Pepsin and trypsin resistant
<i>Paenibacillus</i> sp. Strain E18 [78]	Bifunctional Endo-β-1,4-xylanase/1,3-1,4-β-glucanase	39,000 Da	Opt 6.5-9.0	50	Oat spelt xylan-358 Barley-β-glucan-133	?
<i>C. thermocellum/Geobacillus thermophilus</i> [24]	Chimeric xylanase-arabinofuranosidase	97,000 Da	Opt 6.0	65	?	?

mammals, but has low activity. A variety of xylanases and glucanases are listed. It includes high-activity glucanases and xylanases, bifunctional xylanase/glucanase enzymes where the same catalytic domain hydrolyzes both substrates and chimeric enzymes. The chimeric enzymes can have any combination of catalytic domains, for example, phytase and endoglucanase, xylanase and glucanase, or other combinations of domains. The bifunctional and chimeric enzymes have the advantage of providing multiple activities in one protein, which has the advantage of using a single promoter.

There are a variety of factors that need to be considered when deciding upon promoter(s) and gene(s) for expression in the alimentary tract of animals. Ideally the introduced transgene should impose the minimal metabolic load with minimal interference of existing physiological and biochemical pathways of the host.

Transgene (protein or RNA):

1. Preferably the transformation of interest should be performed by a single molecule. If more than one protein/RNA is involved it may be possible to design the transgene which express multiple transgenes (see below).
2. The transgene must code for a low molecular mass stable protein with a high specific activity, and targeted to the correct compartment.
3. Cryptic posttranslational modification sites are often present in proteins isolated from non-mammalian species and must be removed if activity/stability is affected (i.e., lysostaphin glycosylation).
4. The pH and temperature optima should correspond closely to that of the site of action which would be pH 2–5 if it should be active in the stomach/proventriculus, or pH 5–8 if secreted into the small intestine where it would be active [13, 20, 34].
5. The enzyme must be resistant to proteases: pepsin if expressed in the stomach, or resistant to tryptic enzymes if expressed and active in the small intestine. If the enzyme were expressed in the salivary glands for activity in the small intestine then it would need to be both pepsin and trypsin resistant, stable at low pH, and highly active at neutral pH.
6. Preferably the enzyme should be isolated from the organism with a Generally Recognized As Safe

(GRAS) status and have no significant homology to allergenic, pathogenic, or toxic proteins. It is also preferable that it not exhibit any similarity to known allergens, and it would even be useful to eliminate possible glycosylation sites to avoid the development of an allergenic nature [15].

Regulatory elements:

1. Promoter should be tightly controlled allowing expression of the transgene only in desired tissue at the desired time and at a specific level. Strong ubiquitous promoters commonly used for transgenic work are undesirable as they often result in transgene toxicity and undesirable side effects. An interesting observation in plants was that repetitious use of the same promoter did not lead to transcriptional silencing [54], and this may also apply to mammals.
2. Gene-specific regulatory sequences (5' and 3' UTR, polyA, introns, Kozak start sequence, etc.) should be used to optimize the transcription and translation of transgene.
3. The codon usage should be modified to correspond closely to that of highly expressed proteins within the biosynthetic tissue [9, 28].
4. Insulators and locus control regions (LCR), such as the chicken P-globin locus, should be used to isolate transgene from effect of neighboring chromatin and stabilize the expression level across multiple generations [8].

Multi-gene Expression

The obvious path for the development of transgenic animals is the expression of multiple genes to further enhance the value of an animal. Regulatory agencies undoubtedly will have a preference for the expression by multiple novel genes in an animal produced by crosses of “previously approved” animals with a single transgene to avoid confusion that could result from the interaction of transgenic traits. Whether these will require a further regulatory submission remains to be determined. Since it will be desirable to develop lines of transgenic livestock homozygous for multiple novel traits, a new approach will be needed to eliminate numerous crosses and continued testing for the presence of each gene. The approved transgenes could also be introduced into new

breeds/species by cotransformation, or sequential transformation of established transgenic lines in the same species. Cotransformation could be done with unlinked constructs using different promoters when different ratios of protein products are required. Multiple transgenes can also be housed on a mammalian artificial chromosome. In cases where the same proportion of the protein product is needed, multicistronic constructs with multiple internal *ribosome* entry site (IRES), or even as a single translation product in which individual proteins are separated by cellular protease sites, *could be used* [52].

A proven strategy for multi-transgenic animal production is to use the recent adaptation of the 2A system from foot-and-mouth disease virus [20, 25]. With this approach the open reading frame consists of multiple individual cDNAs separated by sequences encoding 2A and furin cleavage sites. A single complex mRNA is produced and translated into a single complex polypeptide that is spontaneously cleaved into individual exogenous proteins at the 2A sites in the endoplasmic reticulum. This results in equimolar ratios of each of the individual transgenic proteins emerging from the Golgi. This expression system has been demonstrated to produce transgenic animals with up to four transgenes being expressed from a single transcript [86].

Sequential introduction of transgenes would benefit from establishing transgene locus which allows high and stable expression and into which multiple transgenes can be introduced and removed as necessary using site-specific recombination technologies such as Cre/lox, Flp/rtt, ϕ C31 integrase, or Gateway systems [7].

A refinement to further speed up the production of homozygous populations is through the application for flow cytometry [29]. Flow cytometric separation of X and Y chromosome-bearing spermatozoa has been demonstrated to be effective in cattle and pigs and resulted in the birth of healthy offspring of the predetermined gender [68, 69].

Future Directions

Intensive selection by current breeding practices have developed livestock that are growing and producing near their physiological upper limits, and any further improvements may enhance stress leading to disease susceptibility. Reduction in the environmental footprint

of livestock will come through: identification and elimination of bottlenecks in metabolic and physiological pathways, introduction of novel traits that enable the digestion of dietary components previously not digested (e.g., phytate, glycans), through the introduction of novel metabolic pathways for synthesis of essential nutrients (e.g., amino acids), and through enhanced disease resistance. These novel transgenic animals will only be accepted by farmers if they are as robust as the currently available conventional livestock. Furthermore, these new genetically engineered animals will only enter the human food chain once consumers are satisfied that the meat products are absolutely safe, since they have the choice between conventional and transgenic products. Therefore, the central issue is the confidence consumers have in the national regulatory system. With the expected increase in human population and increasing environmental constraints, greater acceptance of transgenic animals is anticipated in line with the acceptance of transgenic plant products.

Bibliography

1. Adams NR, Briegel JR, Ward KA (2002) The impact of a transgene for ovine growth hormone on the performance of two breeds of sheep. *J Anim Sci* 80:2325–2333
2. Bai Y, Wang J, Zhang Z, Shi P, Luo H, Huang H, Feng Y, Yao B (2010a) Extremely acidic β -1,4-glucanase, CelA4, from thermoacidophilic *Alicyclobacillus* sp. A4 with high protease resistance and potential as a pig feed additive. *J Agric Food Chem* 58:1970–1975
3. Bai Y, Wang J, Zhang Z, Shi P, Luo H, Huang H, Luo C, Yao B (2010b) Expression of an extremely acidic β -1,4-glucanase from thermoacidophilic *Alicyclobacillus* sp. A4 in *Pichia pastoris* is improved by truncating the gene sequence. *Microb Cell Fact* 9:33
4. Barkema HW, Green MJ, Bradley AJ, Zadoks RN (2009) Invited review: the role of contagious disease in udder health. *J Dairy Sci* 92:4717–4729
5. Beauchemin KA, Colombatto D, Morgavi DP (2004a) A rationale for the development of feed enzyme products for ruminants. *Can J Anim Sci* 84:23–36
6. Beauchemin KA, Colombatto D, Morgavi DP, Yang WZ, Rode LM (2004b) Mode of action of exogenous cell wall degrading enzymes for ruminants. *Can J Anim Sci* 84:13–22
7. Brown WRA, Lee NCO, Xu Z, Smith MCM (2011) Serine recombinases as tools for genome engineering. *Methods* 53:372–379
8. Bulger M, Groudine M (2010) Enhancers: the abundance and function of regulatory sequences beyond promoters. *Dev Biol* 339:250–257

9. Cannarozzi G, Schraudolph NN, Faty M, von Rohr RP, Friberg MT, Roth AC, Gonnert P, Gonnert G, Barral Y (2010) A role for codon order in translation dynamics. *Cell* 141:355–367
10. Carew R (2010) Ammonia emissions from livestock industries in Canada: feasibility of abatement strategies. *Environ Pollut* 158:2618–2626
11. Cheung QC, Turner PV, Song C, Wu D, Cai HY, Macinnes JL, Li J (2008) Enhanced resistance to bacterial infection in protegrin-1 transgenic mice. *Antimicrob Agents Chemother* 52:1812–1819
12. Cho J, Choi K, Darden T, Reynolds PR, Petitte JN, Shears SB (2006) Avian multiple inositol polyphosphate phosphatase is an active phytase that can be engineered to help ameliorate the planet's "phosphate crisis". *J Biotechnol* 126:248–259
13. Clemens ET, Stevens CE, Southworth M (1975) Sites of organic acid production and pattern of digesta movement in the gastrointestinal tract of swine. *J Nutr* 105:759–768
14. Clements JE, Wall RJ, Narayan O, Hauer D, Schoborg R, Sheffer D, Powell A, Carruth LM, Zink MC, Rexroad CE (1994) Development of transgenic sheep that express the Visna virus envelope gene. *Virology* 200:370–380
15. CODEX 2008 Report of the seventh session of the CodexAd Hoc intergovernmental task force on foods derived from biotechnology. <http://www.codexalimentarius.net/web/archives.jsp?year=08>
16. Cordell D, Drangert JO, White S (2009) The story of phosphorus: global food security and food for thought. *Glob Environ Change* 19:292–305
17. Cowieson AJ, Bedford MR, Selle PH, Ravindran V (2009) Phytate and microbial phytase: implications for endogenous nitrogen losses and nutrient availability. *World's Poult Sci J* 65:401–417
18. Cowieson AJ, Bedford MR, Ravindran V (2010) Interactions between xylanase and glucanase in maize-soy-based diets for broilers. *Brit Poult Sci* 51:246–257
19. Dahlke F, Ribeiro AML, Kessler AM, Lima AR, Maiorka A (2003) Effects of corn particle size and physical form of the diet on the gastrointestinal structures of broiler chickens. *Braz J Poult Sci* 5:61–67
20. D'Apice AJ, Cowan PJ (2008) Gene-modified pigs. *Xenotransplantation* 15:87–90
21. Denning C, Burl S, Ainslie A, Bracken J, Dinnyes A, Fletcher J, King T, Ritchie M, Ritchie WA, Rollo M, de Sousa P, Travers A, Wilmut I, Clark AJ (2001) Deletion of the $\alpha(1,3)$ galactosyl transferase (*GGTA1*) gene and the prion protein (*PrP*) gene in sheep. *Nat Biotechnol* 19:559–562
22. Devlin RH, Sakhrani D, Tymchuk WE, Rise ML, Goh B (2009) Domestication and growth hormone transgenesis cause similar changes in gene expression in coho salmon (*Oncorhynchus kisutch*). *Proc Natl Acad Sci USA* 106:3047–3052
23. Fan CL, Han XY, Xu ZR, Wang LJ, Shi LR (2009) Effects of β -glucanase and xylanase supplementation on gastrointestinal digestive enzyme activities of weaned piglets fed a barley-based diet. *J Anim Physiol Anim Nutr* 93:271–276
24. Fan Z, Wagschal K, Lee CC, Kong Q, Shen KA, Maiti IB, Yuan L (2009) The construction and characterization of two xylan-degrading chimeric enzymes. *Biotechnol Bioeng* 102:684–692
25. Fang J, Qian JJ, Yi S, Harding TC, Tu GH, VanRoey M, Jooss K (2005) Stable antibody expression at therapeutic levels using the 2A peptide. *Nat Biotechnol* 23:584–590
26. Food and Agriculture Organization (2006) World agriculture: towards 2030/2050-interim report. Food and Agriculture Organization, Rome
27. Food and Agriculture Organization (2009) State of food and agriculture – Livestock in the balance. Food and Agriculture Organization, Rome
28. Fredrick K, Ibba M (2010) How the sequence of a gene can tune its translation. *Cell* 141:227–229
29. Garner DL (2006) Flow cytometric sexing of mammalian sperm. *Theriogenology* 65:943–957
30. Gerber PJ, Vellinga TV, Steinfeld H (2010) Issues and options in addressing the environmental consequences of livestock sector's growth. *Meat Sci* 84:244–247
31. Golovan S, Wang G, Zhang J, Forsberg CW (2000) Characterization and overproduction of the *Escherichia coli appA* encoded bifunctional enzyme which exhibits both phytase and acid phosphatase activities. *Can J Microbiol* 46:59–71
32. Golovan SP, Hayes MA, Phillips JP, Forsberg CW (2001) Transgenic mice expressing bacterial phytase as a model for phosphorus pollution control. *Nat Biotechnol* 19:429–433
33. Golovan SP, Meidinger RG, Ajakaiye A, Cottrill M, Wiederkehr MZ, Barney D, Plante C, Pollard J, Fan MZ, Hayes MA, Laursen J, Hjorth JP, Hacker RR, Phillips JP, Forsberg CW (2001) Pigs expressing salivary phytase produce low phosphorus manure. *Nat Biotechnol* 19:741–745
34. Guinotte F, Gautron J, Soumarmon A, Robert JC, Peranzi G, Nys Y (1993) Gastric acid secretion in the chicken: effect of histamine H_2 antagonists and H^+/K^+ -ATPase inhibitors on gastro-intestinal pH and of sexual maturity calcium carbonate level and particle size on proventricular H^+/K^+ + ATPase activity. *Comp Biochem Physiol Comp Physiol* 106:319–327
35. Hall J, Ali S, Surani MA, Hazlewood GP, Clark AJ, Simons JP, Hirst BH, Gilbert HJ (1993) Manipulation of the repertoire of digestive enzymes secreted into the gastrointestinal tract of transgenic mice. *Nat Biotechnol* 11:376–379
36. Hess M, Sczyrba A, Egan R, Kim TW, Chokhawala H, Schroth G, Luo S, Clark DS, Chen F, Zhang T, Mackie RI, Pennacchio LA, Tringe SG, Visel A, Woyke T, Wang Z, Rubin EM (2011) Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science* 331:463–467
37. Huang H, Zhang R, Fu D, Luo J, Li Z, Luo H, Shi P, Yang P, Diao Q, Yao B (2011) Diversity, abundance and characterization of ruminal cysteine phytases suggest their important role in phytate degradation. *Environ Microbiol* 13:747–757
38. Jackson KA, Berg JM, Murray JD, Maga EA (2010) Evaluating the fitness of human lysozyme transgenic dairy goats: growth and reproductive traits. *Transgenic Res* 19:977–986

39. Jerrett SA, Goodge WR (1973) Evidence for amylase in avian salivary glands. *J Morphol* 139:27–45
40. Ji F, Casper DP, Brown PK, Spangler DA, Haydon KD, Pettigrew JE (2008) Effects of dietary supplementation of an enzyme blend on the ileal and fecal digestibility of nutrients in growing pigs. *J Anim Sci* 86:1533–1543
41. Keita D, Heath L, Albina E (2010) Control of African swine fever virus replication by small interfering RNA targeting the A151R and VP72 genes. *Antivir Ther* 15:727–736
42. Kellogg RL, Lander CH, Moffitt DC, Gollehon N (2000) Manure nutrients relative to the capacity of cropland and pastureland to assimilate nutrients: spatial and temporal trends for the United States. <http://www.nrcs.usda.gov/technical/NRI/pubs/mantr.html> 1–140
43. Knowlton KF, Taylor MS, Hill SR, Cobb C, Wilson KF (2007) Manure nutrient excretion by lactating cows fed exogenous phytase and cellulase. *J Dairy Sci* 90:4356–4360
44. Leahy SC, Kelly WJ, Altermann E, Ronimus RS, Yeoman CJ, Pacheco DM, Li D, Kong Z, McTavish S, Sang C, Lambie SC, Janssen PH, Dey D, Attwood GT (2010) The genome sequence of the rumen methanogen *Methanobrevibacter ruminantium* reveals new possibilities for controlling ruminant methane emissions. *PLoS One* 5:e8926
45. Levasseur A, Asther M, Record E (2005) Overproduction and characterization of xylanase B from *Aspergillus niger*. *Can J Microbiol* 51:177–183
46. Lo D, Pursel V, Linton PJ, Sandgren E, Behringer R, Rexroad C, Palmiter RD, Brinster RL (1991) Expression of mouse IgA by transgenic mice, pigs and sheep. *Eur J Immunol* 21:1001–1006
47. Long CR, Tessanne KJ, Golding MC (2010) Applications of RNA interference-based gene silencing in animal agriculture. *Reprod Fertil Dev* 22:47–58
48. Luo H, Yang J, Yang P, Li J, Huang H, Shi P, Bai Y, Wang Y, Fan Y, Yao B (2010) Gene cloning and expression of a new acidic family 7 endo- β -1,3-1,4-glucanase from the acidophilic fungus *Bispora* sp. MEY-1. *Appl Microbiol Biotechnol* 85:1015–1023
49. Lyall J, Irvine RM, Sherman A, McKinley TJ, Néñz A, Purdie A, Outtrim L, Brown IH, Rolleston-Smith G, Sang H, Tiley L (2011) Suppression of Avian influenza transmission in genetically modified chickens. *Science* 331:223–226
50. Maenz DD, Classen HL (1998) Phytase activity in the small intestinal brush border membrane of the chicken. *Poult Sci* 77:557–563
51. Maga EA, Cullor JS, Smith W, Anderson GB, Murray JD (2006) Human lysozyme expressed in the mammary gland of transgenic dairy goats can inhibit the growth of bacteria that cause mastitis and the cold-spoilage of milk. *Foodborne Pathog Dis* 3:384–392
52. Martinez-Salas E (1999) Internal ribosome entry site biology and its use in expression vectors. *Curr Opin Biotechnol* 10:458–464
53. Muller M, Brem G (1994) Transgenic strategies to increase disease resistance in livestock. *Reprod Fertil Dev* 6: 605–613
54. Naqvi S, Farre G, Sanahuja G, Capell T, Zhu C, Christou P (2010) When more is better: multigene engineering in plants. *Trends Plant Sci* 15:48–56
55. Noble MS, Rodriguez-Zas S, Cook JB, Bleck GT, Hurley WL, Wheeler MB (2002) Lactational performance of first-parity transgenic gilts expressing bovine α -lactalbumin in their milk. *J Anim Sci* 80:1090–1096
56. Notter DR (1999) The importance of genetic diversity in livestock populations of the future. *J Anim Sci* 77:61–69
57. Nottle MB, Nagashima H, Verma PJ, Du ZT, Grupen CG, McIlpatrick SM, Ashman RJ, Harding MP, Giannakis C, Wigley PL, Lyons IG, Harrison DT, Luxford BG, Campbell RG, Crawford RJ, Robins AJ (1999) Production and analysis of transgenic pigs containing a metallothionein porcine growth hormone gene construct. In: *Transgenic animals in agriculture*. Commonwealth Agriculture Bureau International, London, pp 145–156
58. NRC (1998) Nutrient requirements of swine. National Academy Press, Washington, DC
59. Nyachoti CM, Arntfield SD, Guenter W, Cenkowski S, Opapeju FO (2006) Effect of micronized pea and enzyme supplementation on nutrient utilization and manure output in growing pigs. *J Anim Sci* 84:2150–2156
60. Ohnishi H, Samuelson LC, Yule DI, Ernst SA, Williams JA (1997) Overexpression of Rab3D enhances regulated amylase secretion from pancreatic acini of transgenic mice. *J Clin Invest* 100:3044–3052
61. Olmedo LA, Samar ME, Avila RE, de Crosa MG, Dettin L (2000) Avian minor salivary glands: an ultrastructural study of the secretory granules in mucous and seromucous cells. *Acta Odontol Latinoam* 13:87–99
62. Peters A, Lebzien P, Meyer U, Borchert U, Bulang M, Flachowsky G (2010) Effect of exogenous fibrolytic enzymes on ruminal fermentation and nutrient digestion in dairy cows. *Arch Anim Nutr* 64:221–237
63. Pursel VG, Pinkert CA, Miller KF, Bolt DJ, Campbell RG, Palmiter RD, Brinster RL, Hammer RE (1989) Genetic engineering of livestock. *Science* 244:1281–1287
64. Pursel VG, Wall RJ, Michell AD, Elsasser TH, Solomon MB, Coleman ME, DeMayo F, Schwartz RJ (1999) Transgenic animals in agriculture. In: *Expression of insulin-like growth factor-1 in skeletal muscle of transgenic swine*. CABI Publishing, New York, pp 131–144
65. Pursel VG, Mitchell AD, Bee G, Elsasser TH, McMurtry JP, Wall RJ, Coleman ME, Schwartz RJ (2004) Growth and tissue accretion rates of swine expressing an insulin-like growth factor I transgene. *Anim Biotechnol* 15:33–45
66. Rajan GH, Morris CA, Carruthers VR, Wilkins RJ, Wheeler TT (1996) The relative abundance of a salivary protein, bSP30, is correlated with susceptibility to bloat in cattle herds selected for high or low bloat susceptibility. *Anim Genet* 27:407–414
67. Rao DECS, Rao KV, Reddy TP, Reddy VD (2009) Molecular characterization, physicochemical properties, known and potential applications of phytases: an overview. *Crit Rev Biotechnol* 29:182–198

68. Rath D, Johnson LA (2008) Application and commercialization of flow cytometrically sex-sorted semen. *Reprod Domest Anim* 43(Suppl 2):338–346
69. Rath D, Ruiz S, Sieg B (2003) Birth of female piglets following intrauterine insemination of a sow using flow cytometrically sexed boar semen. *Vet Rec* 152:400–401
70. Reddy VA, Venu K, Rao DE, Rao KV, Reddy VD (2009) Chimeric gene construct coding for bi-functional enzyme endowed with endoglucanase and phytase activities. *Arch Microbiol* 191:171–175
71. Rees WD, Hay SM (1995) The biosynthesis of threonine by mammalian cells: expression of a complete bacterial biosynthetic pathway in an animal cell. *Biochem J* 309(Pt 3):999–1007
72. Rich KM, Perry BD (2010) The economic and poverty impacts of animal diseases in developing countries: New roles, new demands for economics and epidemiology. *Prev Vet Med* 97:1–8
73. Richt JA, Kasinathan P, Hamir AN, Castilla J, Sathiyaseelan T, Vargas F, Sathiyaseelan J, Wu H, Matsushita H, Koster J, Kato S, Ishida I, Soto C, Robl JM, Kuroiwa Y (2007) Production of cattle lacking prion protein. *Nat Biotechnol* 25:132–138
74. Romano PR, Wang J, O'Keefe RJ, Puzas JE, Rosier RN, Reynolds PR (1998) HIPER1, a phosphatase of the endoplasmic reticulum with a role in chondrocyte maturation. *J Cell Sci* 111:803–813
75. Samar ME, Avila RE, Esteban FJ, Olmedo L, Dettin L, Massone A, Pedrosa JA, Peinado MA (2002) Histochemical and ultrastructural study of the chicken salivary palatine glands. *Acta Histochem* 104:199–207
76. Schröder JJ, Smit AL, Cordell D, Rosemarin A (2011) Improved phosphorus use efficiency in agriculture: a key requirement for its sustainable use. *Chemosphere* 84(6):822–831
77. Seo HW, Kim TM, Choi JW, Han BK, Song G, Han JY (2010) Evaluation of combinatorial cis-regulatory elements for stable gene expression in chicken cells. *BMC Biotechnol* 10:69
78. Shi P, Tian J, Yuan T, Liu X, Huang H, Bai Y, Yang P, Chen X, Wu N, Yao B (2010) *Paenibacillus* sp. strain E18 bifunctional xylanase-glucanase with a single catalytic domain. *Appl Environ Microbiol* 76:3620–3624
79. Simojoki H, Orro T, Taponen S, Pyorala S (2009) Host response in bovine mastitis experimentally induced with *Staphylococcus chromogenes*. *Vet Microbiol* 134:95–99
80. Simojoki H, Hyvonen P, Orro T, Pyorala S (2010) High concentration of human lactoferrin in milk of rhLf-transgenic cows relieves signs of bovine experimental *Staphylococcus chromogenes* intramammary infection. *Vet Immunol Immunopathol* 136:265–271
81. Skinner JT, Bauer S, Young V, Pauling G, Wilson J (2010) An economic analysis of the impact of subclinical (mild) necrotic enteritis in broiler chickens. *Avian Dis* 54:1237–1240
82. Smith TC, Harper AL, Nair R, Wardyn SE, Hanson BM, Ferguson DD, Dressler AE (2011) Emerging swine zoonoses. *Vector Borne Zoonotic Dis* 11:1–10
83. Steinfeld H, Gerber P (2010) Livestock production and the global environment: consume less or produce better? *Proc Natl Acad Sci USA* 107:18237–18238
84. Steinfeld H, Gerber P, Wassenaar T, Castel V, Rosales M, de Hann C (2006) Livestock's long shadow – Environmental issues and options. FAO, Rome
85. Suzuki M, Fujimoto W, Goto M, Morimatsu M, Syuto B, Iwanaga T (2002) Cellular expression of gut chitinase mRNA in the gastrointestinal tract of mice and chickens. *J Histochem Cytochem* 50:1081–1089
86. Szymczak AL, Workman CJ, Wang Y, Vignali KM, Dilioglou S, Vanin EF, Vignali DA (2004) Correction of multi-gene deficiency in vivo using a single 'self-cleaving' 2A peptide-based retroviral vector. *Nat Biotechnol* 22:589–594
87. Tilman D, Fargione J, Wolff B, D'Antonio C, Dobson A, Howarth R, Schindler D, Schlesinger WH, Simberloff D, Swackhamer D (2001) Forecasting agriculturally driven global environmental changes. *Science* 292:281–284
88. Wall RJ, Powell AM, Paape MJ, Kerr DE, Bannerman DD, Pursel VG, Wells K, Talbot N, Hawk HW (2005) Genetically enhanced cows resist intramammary *Staphylococcus aureus* infection. *Nat Biotechnol* 23:445–451
89. Wen TN, Chen JL, Lee SH, Yang NS, Shyr LF (2005) A truncated *Fibrobacter succinogenes* 1,3-1,4- β -D-glucanase with improved enzymatic activity and thermotolerance. *Biochemistry* 44: 9197–9205
90. Wheeler MB, Bleck GT, Donovan SM (2001) Transgenic alteration of sow milk to improve piglet growth and health. *Reprod Suppl* 58:313–324
91. Wheeler TT, Hood K, Oden K, McCracken J, Morris CA (2003) Bovine parotid secretory protein: structure, expression and relatedness to other BPI (bactericidal/permeability-increasing protein)-like proteins. *Biochem Soc Trans* 31:781–784
92. Wise TG, Schafer DS, Lowenthal JW, Doran TJ (2008) The use of RNAi and transgenics to develop viral disease resistant livestock. *Dev Biol (Basel)* 132:377–382
93. Xue GP, Denman SE, Glassop D, Johnson JS, Dierens LM, Gobius KS, Aylward JH (1995) Modification of a xylanase cDNA isolated from an anaerobic fungus *Neocallimastix patriciarum* for high-level expression in *Escherichia coli*. *J Biotechnol* 38:269–277
94. Yang W-J, Matsuda Y, Sano S, Masutani H, Nakagawa H (1991) Purification and characterization of phytase from rat intestinal mucosa. *Biochim Biophys Acta* 1075:75–82
95. Yin HF, Fan BL, Yang B, Liu YF, Luo J, Tian XH, Li N (2006) Cloning of pig parotid secretory protein gene upstream promoter and the establishment of a transgenic mouse model expressing bacterial phytase for agricultural phosphorus pollution control. *J Anim Sci* 84:513–519
96. Zhang JX, Meidinger R, Forsberg CW, Krell PJ, Phillips JP (1999) Expression and processing of a bacterial endoglucanase in transgenic mice. *Arch Biochem Biophys* 367:317–321
97. Zhang F, Shi P, Bai Y, Luo H, Yuan T, Huang H, Yang P, Miao L, Yao B (2011) An acid and highly thermostable xylanase from *Phialophora* sp. G5. *Appl Microbiol Biotechnol* 89:1851–1858

Transgenic Livestock, Enhanced Nutritional Quality in

GÖTZ LAIBLE

AgResearch, Ruakura Research Centre, Hamilton,
New Zealand

Article Outline

Glossary

Definition of the Subject and Its Importance

Introduction

Milk with Enhanced Nutritional Qualities

Meat with Enhanced Nutritional Qualities

Safer Food from Livestock with Enhanced Health

Characteristics

Future Directions

Bibliography

Glossary

Antimicrobial A substance with the ability to kill microbes.

Bacteriolytic Causing the lysis of bacteria.

Bacteriostatic Term to describe a substance that inhibits bacterial growth.

Casein micelle Protein particle consisting of aggregated casein proteins in colloidal suspension.

Circadian rhythm A daily cycle controlling the biological processes in living organisms.

Dominant-negative molecule A mutant molecule capable of interacting with the wild-type form to produce an inactive complex.

Endogenous Internally derived or synthesized.

Endopeptidase Proteolytic enzyme that can only break peptide bonds within the molecule but not at the termini.

Gastrointestinal Term referring to the digestive tract.

Genome Entirety of an organism's hereditary information.

Glycomacropeptide κ -casein-derived peptide released by rennet during cheese manufacture.

Homologous recombination Exchange of genetic material between two DNA fragments by crossing over in a region with sequence homology.

In vitro Experimental procedure conducted artificially.

Knockdown Attenuation of gene function usually resulting in diminished amount of synthesis of the protein that the gene encodes.

Knockin Integration of a new gene which replaces an endogenous gene.

Knockout Functional disruption of a specific gene of an organism commonly achieved by a partial or complete deletion of the gene sequence.

Pathogen/pathogenic Infectious agent/disease causing.

Phenotype/phenotypic A measurable characteristic of an animal such as hair color growth rate, or degree of carcass marbling. These traits are the product of genetics and the environment.

Phenylketonuria Genetic disorder in which the essential amino acid phenylalanine cannot be metabolized potentially causing mental retardation.

Pleiotropic effects The phenomena of a single gene having influences on multiple traits.

Prion An infectious protein particle.

Promoter A regulatory DNA sequence that controls the transcription of a particular gene.

RNA interference A sequence-specific gene-silencing process in which double-stranded RNAs trigger the destruction of specific RNAs.

Somatic cell count Quantification of somatic cells found in milk as an indicator for the quality of the milk.

Transgene An exogenous gene introduced into the genome of another organism.

Transgenic An animal plant or microbe whose genetic material has been altered using an artificial process.

Unsaturated (saturated) fatty acids Fatty acid molecules with (no) double bonds.

Whey Liquid that remains after separation of the solid fraction when cheese is made.

Definition of the Subject and Its Importance

Since the domestication of animals several millennia ago, which enabled the transition from hunter-gatherer to farming communities, humans have shaped livestock according to specific needs for food, both in quantity and quality. Industrialization and rapid population growths prompted immense changes to farming systems, human lifestyle, and food choices, and is predicted to result in drastic climate changes within

a relative short time period. Thus, current food production systems face new challenges in this rapidly changing world, in particular demands for safe and healthier food to address modern health concerns and the desire for longevity. In the past, traditional breeding and selection schemes have been invaluable for the incremental improvement of food animals. However, the process is slow and undirected and is not well suited for enhancing specific nutritional characteristics. The relatively new transgenic technology offers the potential for enhancing existing or introducing entirely novel characteristics at unprecedented rates and magnitudes and could provide substantial benefits for consumers in the form of healthier and safer food. Nevertheless, the complexity of food production traits should not be underestimated. The accurate modification of the appropriate gene(s) to generate desired phenotypes able to deliver enhanced food products remains challenging and requires careful testing and evaluation. Despite the unrivaled opportunities for innovative functional foods and the progress realized with transgenic crops, lack of acceptance due to concerns associated with the technological novelty of transgenic livestock has so far greatly limited even the research-driven evaluation of new concepts aimed at food applications.

Introduction

Livestock animals are farmed for human food production and are a primary source of dietary proteins, lipids, carbohydrates, vitamins, and minerals from milk and meat. The development of efficient farming production systems has played a crucial role to satisfy the elementary requirement for a stable food source to meet the daily nutritional needs of a multibillion human population. While milk and meat represent in general, a high-quality food source for humans, the effects of modernization and associated substantial changes to the farming systems and lifestyle of people create new challenges for providing a healthy diet of high nutritional quality. The increased understanding of how human nutritional requirements relate to health and longevity provides new opportunities to optimize the nutritional quality of food products and thus improve human health and well-being. Animal fat, which is an ingredient of all animal food products,

can serve as an example to illustrate this correlation. They are relatively high in saturated fats (which have been strongly associated with cardiovascular and coronary heart disease), but low in beneficial, unsaturated fats. Thus, higher levels of unsaturated fats and lower levels of saturated fats would improve the nutritional quality of animal-derived food products. Efforts to increase the ratio of unsaturated to saturated fat have received considerable attention. However, attempts to alter the composition of milk and meat to improve its nutritional quality by conventional breeding and selection strategies have proved to be difficult. Traditional breeding and marker-assisted selection schemes are limited to the random combination of some but not all superior allelic gene variants existing within the gene pool of the livestock species. In contrast, transgenic animal technology allows for a more targeted approach with the prospect to introduce specific genes which are known to impact on defined phenotypic traits to enhance existing characteristics. Unlike traditional breeding and selection, the genetic improvement of livestock by genetic engineering is not restricted by the species barrier and can utilize the gene pool of other species to introduce entirely novel and unique characteristics. Moreover, transgenic technology is a versatile platform technology that can employ additive strategies to introduce a new gene function (gain of function), delete gene functions (knockout, loss of function), replace an endogenous gene function with a different one (knockin, exchange of function), or precisely control when and where the genetic alteration is applied (inducible gene expression and conditional knockout). Thus, it holds great promise as a new tool that can deliver solutions for many of the problems concerning the efficient production of high-quality foods from animals.

Despite mounting lifestyle-related health problems, food security, and environmental pressures which need to be urgently addressed, negative public perception, ethical concerns, regulatory uncertainties, and an associated reluctance to fund this line of research have been major factors that have so far greatly limited the international research efforts to evaluate the potential of the transgenic technology for enhancing food production of livestock. While transgenic plants have been commercialized since 1996 [1], transgenic livestock applications for

food production have so far been restricted to a few proof-of-concept studies that aim to improve the nutritional quality of milk and meat and will be outlined in detail in the following sections.

Milk with Enhanced Nutritional Qualities

Milk is an important food with high protein content. The protein fraction determines many of the functional properties and provides opportunities for nutritional enhancement through additional health benefits, increased food safety, and improved processing properties for dairy food manufacture.

One approach has been the introduction of new antimicrobial properties into milk, to provide passive immunity to consumers. Lysozyme (LZ) is a naturally occurring antimicrobial protein, present in milk, saliva, and tears of mammals, where it forms part of the defense system against bacterial infections. However, it is found at three orders of magnitude higher concentration in human milk as compared to dairy milk. The introduction of an expression construct for human LZ resulted in LZ levels in the milk of the transgenic goats equivalent to about 68% of the level found in human milk [2]. The consumption of the human LZ-containing goat milk by pigs exerted beneficial effects and improved their gastrointestinal health [3, 4]. Although the milk has not been tested in humans, the promising results suggest similar gastrointestinal benefits could be possible for humans. In addition, the elevated LZ levels increased general food safety properties of the milk by reducing the bacterial growth that causes cold-spoilage thus prolonging its shelf life. Following a similar strategy, recombinant human lactoferrin (LF) has been overexpressed in the milk of dairy cattle [5]. LF is a minor milk protein with antimicrobial and anti-inflammatory properties, which can support the innate defense mechanisms. Although bovine LF has similar functions, it is present in milk at very low levels. Its human counterpart has been fully adapted to human requirements during evolution and is expected to be better suited for human health applications, in particular, at elevated levels. The transgenic cows were shown to produce milk that contains human LF at concentrations much higher than the natural level of bovine LF. These new characteristics indicate that this functional food might offer new health benefits such as increased

protection against infections and healthier intestinal microflora. A conceptually related approach aims to boost the immune system through food-mediated delivery of neutralizing monoclonal antibodies to target specific pathogenic microorganisms of the digestive tract. Thus, consumption of antibody-enriched milk could provide instant protection from infections which severely affect those with increased susceptibility, including the elderly, infants, and patients with illnesses. Proof-of-concept studies in mice successfully demonstrated the potential of this approach. The expression of high antibody levels in milk was able to confer complete protection for the suckling young when challenged with an otherwise lethal dose of virus [6]. Whether the health-enhancing attributes of the functional proteins described above can indeed be realized as food-mediated health benefits in humans is presently unknown and still needs to be assessed.

Although milk from dairy cows is a very common human food, it is known to trigger allergic reactions in some people, a phenomenon that is particularly problematic in infants. Not present in human milk, β -lactoglobulin has been identified as a major whey protein in the milk of dairy species and is thought to be a major allergen in the milk of cows. Reduction of this allergenic protein in dairy milk would make it more amenable for susceptible individuals. This could be achieved by either the disruption of the β -lactoglobulin gene using homologous recombination, often referred to as knockout, or the specific knock-down of its expression by RNA interference (RNAi). However, gene knockouts, a standard approach for the functional characterization of genes in mice, proved to be difficult in farm animals due to low recombination efficiencies in primary livestock cells. Thus RNAi technology might provide a more straightforward approach to unravel the role of β -lactoglobulin and reduce the allergenicity of cow's milk.

Altering the milk composition by altering the levels and ratios of the main endogenous milk proteins has been suggested to enhance the nutritional quality of milk and its processing characteristics, including increased heat stability, calcium content, and cheese manufacture [7, 8]. A drastic change of milk composition has been achieved recently with the introduction of additional β - and κ -casein genes in transgenic cattle [9]. The milk derived from these cows showed a two- to

threefold increase in κ -casein and a minor change for β -casein. These changes affected the physical appearance of the milk, which showed a distinctive color change from ordinary white to yellow for the modified milk [10]. Because κ -casein is located on the surface of the casein micelles, any increase in κ -casein is expected to reduce the size of the casein micelles [11]. The reduced size of the micelles affects the light scattering properties and is thought to be the reason for the observed color change in the high κ -casein milk [10]. Although the effects on the processing characteristics of this novel milk still remain to be determined, trial cheese production resulted in cheese with increased levels of essential amino acids and beneficial minerals and thus improved nutritional quality.

Furthermore, due to the higher κ -casein content, this speciality milk is an improved source of κ -casein-derived bioactive peptides [12] such as glycomacropeptide (GMP) which has been associated with a wide range of health-promoting activities including protection against toxins, bacteria, and viruses, suppression of gastrointestinal inflammation, and modulation of the immune system [13]. GMP is one of the few naturally occurring proteins that lacks the amino acid phenylalanine, which makes it a safe source of dietary protein for people suffering from the genetic disorder phenylketonuria (PKU). They cannot metabolize phenylalanine, and normally need to avoid high-protein foods such as dairy products to prevent the detrimental accumulation of phenylalanine metabolites.

An extension to this, highlighting the unique ability of the transgenic technology to provide novel foods tailored for specific dietary requirements, is the development of transgenic rabbits that produce a milk protein suitable as dietary replacement for PKU sufferers [14].

The carbohydrate component of dairy milk has been another target for modifying milk composition to improve health attributes. The milk sugar lactose causes intestinal disorders in lactose-intolerant people who lack adequate intestinal lactose-hydrolyzing enzyme activity to sufficiently digest lactose after milk consumption [15]. Two strategies to reduce the milk sugar lactose content were evaluated in transgenic mouse models which could provide an elegant alternative to expensive postharvest milk processing. The complete disruption of the expression of α -lactalbumin, an essential component of the

lactose synthetase complex, resulted in lactose-free milk [16]. However, because lactose is the main osmotic regulator of milk secretion, it strongly affected production and secretion of the milk, which was highly concentrated and the transgenic mice were unable to sustain lactation. In comparison to the knockout strategy, knockdown of α -lactalbumin expression through RNAi could offer better control, to achieve an acceptable reduction of the lactose and water content of milk, without impacting on its vital attributes. Thus, it might provide an opportunity to reduce the lactose content and at the same time significantly lower transportation costs of liquid milk.

A more successful approach addressing lactose intolerance, which cleverly dissected the issues of osmolarity and lactose levels, targeted the expression of a mammalian lactose hydrolyzing enzyme to the mammary gland [17]. Milk lactose was reduced by 50–85% without affecting osmolarity, due to the conversion of the produced lactose into the osmotically active monosaccharides, glucose, and galactose. Although the transgenic mouse results are promising, and it was speculated that a similar reduction of lactose in bovine milk could ameliorate lactose intolerance, they can only provide an indication due to substantial species-specific differences in milk composition. So far, this approach has not been transferred to livestock and still awaits verification of its merit in dairy cows.

Beside protein and sugar, fat is another important nutritional component of milk which has a significant impact on human health and provides opportunities for the improvement of dairy products. Animal fats are very rich in saturated fatty acids (SFA) which are considered “unhealthy” and have been associated with cardiovascular and coronary heart disease. A major focus is therefore to decrease the unhealthy fats in favor of the healthier unsaturated fats. This would have the added benefit of decreasing the hardness of milk fat resulting in softer butter, with improved spreadability. As the majority of the long-chain fatty acids in milk fat and in particular unsaturated fatty acids (UFAs) originate from food sources, an effective approach to improve milk fat composition might be to tailor the diet of dairy cows through the use of particular forage species or feed additives [18, 19]. Transgenic technology on the other hand provides the ability of a more directed approach, by modulating endogenous

enzymes involved in lipid metabolism. The mammary gland-specific expression in transgenic goats of the rat stearoyl-CoA desaturase (SCD), an enzyme involved in converting SFAs into mono-UFAs, resulted in an increase of mono-UFAs and decreased levels of medium chain SFAs [20]. However, this beneficial change in the milk fat composition was only transient in this particular study. It was most prominent in early lactation which was thought to be a consequence of high instability of the SCD-encoding mRNA and the resulting low expression levels.

Another concept has been to lower the total fat content by disrupting acetyl-coenzyme A carboxylase to prevent *de novo* fatty acid synthesis in the mammary gland, which accounts for about 50% of the milk fat [7]. While low-fat liquid milk is desirable from a consumer perspective, for the cow it could reduce the feed energy requirements and improve body condition which is likely to translate into lower milk production costs and improved conception rates during early lactation.

A more extreme variation of the concept of adjusting the activity of endogenous enzymes involved in lipid metabolism is the introduction of novel enzymatic activities normally not found in mammals. This can provide livestock animals with the ability to endogenously synthesize essential poly-UFAs (PUFA) and transform their food products into enriched sources. As a result, food products derived from these animals could offer a range of additional PUFA-specific health benefits such as reduced risk for the development of coronary heart disease and a healthy immune system [21]. This strategy was first tested in transgenic mice with the introduction of the fat-1 gene encoding the *Caenorhabditis elegans* n-3 fatty acid desaturase, an enzyme activity that is absent in vertebrates and renders transgenic animals capable of producing omega-3 FAs through conversion of dietary derived n-6 FAs. Omega-3 is a class of beneficial n-3 PUFAs that is associated with a lower risk of morbidity and mortality from atherosclerosis and coronary heart disease. Milk-specific expression of the desaturase in transgenic mice resulted in elevated levels of long-chain n-3 PUFAs and a concomitant decrease in n-6 PUFAs, although this was most pronounced in phospholipids which

are a minor fraction of milk fat, and to a lesser degree in the milk triacylglycerides [22]. Pups consuming the n-3 PUFA enriched milk accumulated the n-3 PUFA docohexaenoic acid in their brains, a compound which has been associated with cognitive performance [23]. The constitutive expression using a humanized fat-1 gene resulted in essentially similar qualitative changes of the n-3 and n-6 PUFAs in milk but at a higher magnitude [24]. So far, the concept has been transferred to pigs but not to any of the dairy species. Unfortunately, changes to the milk fat composition were not investigated in these pigs because the main focus for the study was on the fat content of meat and will be discussed in the next section.

Meat with Enhanced Nutritional Qualities

The introduction of a novel FA desaturase activity to genetically complement mammals with the ability to endogenously synthesize essential PUFA and improve the nutritional characteristics of meat was pioneered in a transgenic mouse model mentioned in the previous section [24]. Mice were engineered by integrating a humanized version of the fat-1 gene into their genomes. Expression of the *Caenorhabditis elegans* n-3 FA desaturase in these transgenic mice resulted in muscles that were highly enriched in the beneficial omega-3 FAs. While all organs and tissues of the transgenic mice showed a marked reduction of the n-6 to n-3 FA ratio, the effect was greatest for skeletal muscle where the ratio dropped from a value of ~50 for conventional mice to close to 1 for transgenic mice. Considering the health concerns associated with the typically high n-6 to n-3 ratio of Western diets, consumption of food products from omega-3 enriched livestock may provide an opportunity to improve diets. When the approach was recently applied to pigs, the constitutive expression of the fat-1 transgene essentially replicated these results with tail samples from the transgenic pigs found to be highly enriched in omega-3 FAs, in particular for two of the most potent n-3 FAs mainly found in fish [25]. Although, the nutritional consequences still need to be investigated in greater detail, transgenic food animals enriched in omega-3 FAs may provide a more economical, safe, and sustainable means to

fortify meat than the current practice of feeding animals with fish meal and satisfy the growing demand for omega-3 fatty acids in human nutrition [26].

Following a similar concept, pigs were engineered for the endogenous production of the essential PUFA linoleic acid by introducing the FAD2 gene encoding the $\Delta 12$ FA desaturase from spinach [27]. Expression of the plant desaturase was selectively targeted to adipocytes and resulted in transgenic pigs showing a 20% increase in linoleic acid content of their white adipose tissue. These results further substantiate the prospect for improving the fatty acid composition of domestic animals through transgenic technology. Meat produced by transgenic pigs could ultimately provide an alternative source for essential FAs, which may ameliorate lifestyle-related health concerns.

Early transgenic livestock studies were mainly focused on modifying body composition for enhanced meat production by stimulating muscle growth via the introduction of genes for growth factors such as growth hormone (GH) or insulin-like growth factor I (IGF-I), a concept that has been highly successful in proof-of-concept studies in transgenic mice [28–31]. Although pigs responded with remarkable growth enhancement following the administration of exogenous GH [32], the initial studies with transgenic GH pigs [33] and sheep [34, 35] were disappointing. The transgenic animals achieved only slightly increased growth rates and were plagued by a range of deleterious side effects due to high systemic GH levels. These were a consequence of poor regulatory control of the transgene activity [33, 36]. Targeting the transgene expression to skeletal muscle [37] or applying conditional expression strategies with the ability to switch it on or off [38] essentially overcame the problem of generating adverse health effects in the transgenic animals and resulted in more desirable effects on growth rate and body composition. Pigs engineered for the tissue-specific expression of human IGF-I in skeletal muscle produced approximately 10% more carcass lean tissue and 20% less total carcass fat. Unexpectedly, the effects showed a strong gender bias with the body composition of transgenic males remaining comparable to conventional boars [37]. Transgenic sheep overexpressing ovine GH, controlled by a zinc-inducible promoter,

grew significantly faster, and had a leaner body composition [39]. However, the transgenic sheep had higher parasite fecal egg counts, a potential indication of a compromised immune system.

The modest success of improving growth characteristics in sheep and pigs with growth factor-encoding transgenes is in complete contrast to what has been achieved with GH-enhanced transgenic fish using all piscine DNA constructs [40–43]. In the majority of fish species, the GH transgene resulted in dramatic growth acceleration with up to 35-fold increased growth rates in transgenic loach and salmonids and produced fish that reach double the normal body size in half the normal time [44]. However, some of these growth-enhanced fish showed undesirable pleiotropic effects such as altered skin color, modified skull shape, decreased fertility, and decreased viability. Intriguingly, the astounding growth phenotypes were only achievable in wild fish and could not be replicated in domesticated fish [45]. This may indicate that there is a biological ceiling for further growth enhancement by GH in highly selected domesticated animals, which could explain why the approach achieved only slightly accelerated growth rates in transgenic GH pigs and sheep. Thus, for livestock, the approach, at least in its current form, appears unlikely to deliver on its promise to substantially increase meat production which has to be achieved without adversely impacting on the health of the transgenic animals. The only beneficial outcome realized was some nutritional enhancement through the production of leaner meat, which may provide new leads to further improve the nutritional quality of meat.

A promising alternative strategy to boost growth performance with the prospect of greater control might be the direct manipulation of key regulators of skeletal muscle development. Myostatin is a crucial negative regulator of muscle growth, whose function was revealed with a mouse knockout model which showed two to three times the muscle mass of wild-type mice [46]. These mice closely resembled the double-muscling phenotype of some cattle breeds characterized by a 20% increase in muscle mass associated with a reduction in fat tissue. Analysis of the myostatin gene revealed that these breeds are carriers of natural mutations that result in loss of myostatin function as the underlying cause for the double-muscle phenotype in cattle [47, 48]. The bulkier conformation of these animals is, however,

causing major calving difficulties in these breeds, which are associated with significant welfare concerns. Transgenic technology provides the possibility to avoid these concerns by restricting the effects on the myostatin pathway to only postnatal stages of major muscle growth. The feasibility of such an approach was already demonstrated with a conditional myostatin knockout in mice resulting in postnatal disruption of the myostatin gene and comparable increase in muscle mass to a constitutive knockout [49]. However, additive strategies to interfere with the myostatin pathway including the expression of dominant-negative molecules [50, 51] or RNAi [52, 53] might offer even more flexibility with the ability to control the degree of increased muscle mass. Using site-specific recombination techniques, transgenic mice were generated with a muscle-specific expression cassette for a dominant-negative myostatin pro-domain integrated into the male-specific Y-chromosome [54]. Males of these lines showed a 5–20% increase in skeletal muscle mass; because females do not have a Y-chromosome, all females were non-transgenic and not affected in their growth characteristics. Combined with a postnatal or inducible expression strategy, interference with myostatin function targeted to males only could improve the efficiency of current cattle production systems with the concomitant production of bulls with superior meat production ability and elite dairy cows.

Safer Food from Livestock with Enhanced Health Characteristics

Any improvement to the health status of food animals has an immediate positive effect on food safety, which ultimately will lead to safer food products of superior quality. Moreover, healthier livestock would deliver a whole range of additional beneficial attributes, including improved animal welfare, reduced reliance on animal remedies, reduced risk for disease transmission to humans, and improved reproductive performance and production. While conventional breeding programs with the aim of reducing susceptibility of livestock to pests and diseases have not been very successful, the prospect of complementing traditional disease control measures with transgenic technology is particularly appealing, because it provides new, targeted strategies for improved disease control and animal health [55, 56].

One of the most costly diseases in agriculture is a bacterial infection of the mammary gland known as mastitis. The disease causes a significant reduction of milk yields and renders the milk produced by infected cows unsuitable and unsafe for human consumption. In its nonclinical appearance, the main indicator is a high somatic cell count in milk. Mastitis, however, has the potential to severely affect the health of infected animals and can cause death or require euthanasia on animal welfare grounds. In dairy cattle, about one third of clinical mastitis cases are caused by *Staphylococcus aureus* infection, a pathogen which, due to its ability of intracellular survival, is particularly difficult to control with conventional antibiotic therapies. Transgenic technology can provide alternative mastitis prevention strategies and in one such approach, the expression of a bacteriolytic enzyme in mammary gland cells, has already been evaluated.

Lysostaphin is a bacterial enzyme with endopeptidase activity that cleaves crucial cell wall components of staphylococci resulting in the lysis of the bacteria. Its potential as an effective antimicrobial agent for the treatment of mastitis was first demonstrated in a mouse model where the mammary-specific expression of lysostaphin conferred a protective effect against *Staphylococcus aureus* infections [57]. Recently, the concept has been successfully extended to cattle. The transgenic cows, which produced lysostaphin in their milk, recapitulated the mouse results and showed a high degree of protection when challenged with the pathogen [58].

Two other proteins with antimicrobial properties, LF and LZ have long been suggested as strong candidates to confer resistance of mastitis in cattle [59]. Although conceptually similar to the lysostaphin approach, cows overexpressing the antimicrobial protein human LF in their milk appeared to gain no improved protection against mastitis when challenged with an *Escherichia coli* strain isolated from cows with clinical mastitis [60]. This was rather unexpected, considering that these transgenic cows produced much greater levels of human LF in milk compared to the milk enriched with lysostaphin. Serving as a model for dairy cattle, human LZ has been expressed in the mammary gland of goats [2]. The potent bacteriostatic properties of the milk produced by these goats against two mastitis causing bacteria strongly suggests that human LZ might be another excellent candidate for conferring resistance to mastitis [61].

Particularly, in light of the human LF results, direct experimental validation of the protective potential will be required to determine the effectiveness of human LZ in preventing mastitis.

A different concept to introduce disease-resistant properties into livestock attempts to strengthen the immune system through the expression of pathogen-specific antibodies and thus providing instant immunity without prior exposure to this particular pathogen. The first studies in livestock involved the expression of mouse monoclonal antibodies in transgenic rabbits, sheep, and pigs but were met with little success because of unexpected problems to express fully functional antibodies [62, 63].

An extension of this approach has been the targeted production of viral-neutralizing monoclonal antibodies in milk, to provide passive immunity and protection from viral infections, which can severely affect suckling neonates. A mouse model expressing a coronavirus specific monoclonal antibody in milk demonstrated that the sustained production of high antibody levels throughout lactation could successfully neutralize the viral infectivity [64]. Moreover, in a similar study, suckling young were challenged with a lethal dose of murine hepatitis virus (MHV). Ingestion of MHV-specific antibody-enriched milk produced by the transgenic mice induced full protection against the viral infection in the pups [6].

The disruption of the virus entry mechanism proved to be an equally potent strategy. Transgenic mice engineered for the expression of a soluble form of a porcine herpes virus receptor, gained effectively full resistance against pseudorabies virus (PRV) infections [65]. Furthermore, the transgenic mice displayed superior protection levels compared to mice vaccinated with an attenuated PRV strain, demonstrating that this transgenic approach may offer superior efficiency for the control of pseudorabies in livestock [66].

The introduction of a specific disease resistance gene MX1 from mouse, known to confer resistance to influenza viruses in mice was another approach evaluated for its potential to protect livestock animals from viral infections. Similar to other pioneering studies prior to the mid-1990s, difficulties to adequately control Mx1 expression in transgenic pigs hampered the approach, which ultimately failed to achieve viral

protection in pigs [67]. For now, the real potential of this approach remains unknown but the identification of antiviral Mx alleles from livestock species [68–70] and new tools for the improved control of transgene expression levels, such as the recently developed autoregulative tetracycline-responsive expression cassette [71], may encourage a reevaluation of MX alleles for their potential to confer disease resistance in domesticated animals.

Rather than attempting to provide additional protective attributes, transgenic technology also enables the direct targeting of endogenous genes implicated in disease pathways. Such a concept has been applied to produce livestock animals which are resistant to the fatal neurodegenerative prion diseases or transmissible spongiform encephalopathies. This family of diseases, which include scrapie and the “mad cow disease” bovine spongiform encephalopathy, or BSE for short, is characterized by an accumulation of a misfolded isoform of the cellular prion protein (PrP) that acts as a novel infectious agent. Proof-of-concept studies in the mouse [72, 73], either introduction of mutated prion protein genes [74], gene knockout [75–77], or RNAi-mediated knockdown of PrP expression [78] demonstrated the possibility to produce animals that are resistant to prion diseases. Resistant livestock, certain to be free of such diseases would eliminate the risk for potential transmission of the disease to humans and provide additional safeguards for the food chain. However, some mouse studies raised concerns that the loss of the normal cellular function of PrP may adversely affect the animals by interfering with the circadian rhythm [79], synaptic functions [80, 81], learning [82], or loss of movement coordination [83]. When these findings were further investigated, the underlying molecular cause was not the loss of PrP function but an interference with the expression pattern of an adjacent gene as a result of only partially deleting the PrP locus [84–86].

Initial studies in sheep [75] and goats [77] indicated that animals heterozygous for the PrP knockout display, at least in young animals, a normal phenotype. These findings could now be confirmed with homozygous knockout cattle with deletions of the entire PrP gene [87]. In vitro assays with brain tissue homogenates derived from these PrP deficient cattle demonstrated resistance to prion propagation while the cattle were apparently normal in all aspects analyzed.

The recent development of RNAi technology, based on the expression of a short RNA molecule that can interfere with the activity of a specific target gene offers new opportunities to directly attack pathogens by disrupting their lifecycles. It is particularly attractive as an antiviral strategy, where RNAi can be used to target viral transcripts and suppress viral infection [88]. Because RNAi is a sequence-specific process, it will only affect the virus but not the host animal. Validated by the successful application of RNAi to control viral diseases in mouse model systems [89–91], the concept is applicable to a wide range of significant pathogens such as the RNA viruses that constitute over two thirds of the Office International Des Epizooties (OIE) list-A pathogens [88]. Recent outbreaks of viral diseases in farmed animals with immense socioeconomic and animal welfare costs such as foot and mouth disease or with the risk to cross the species barrier into humans such as Severe Acute Respiratory Syndrome (SARS), avian influenza and swine flu highlight the importance for improving current intervention strategies. Several studies are presently undertaken to evaluate the effectiveness of RNAi-mediated viral resistance in domesticated animals including chickens, pigs, cattle, sheep, and horses.

Future Directions

Transgenic livestock technology has made tremendous progress since its humble beginnings in 1985 [92]. Poor control of the transgene activity presented a main problem in early studies. Today, it is possible to integrate a transgene into a specific site of the genome and provide accurate control from the endogenous regulatory sequence elements. In addition, molecular switches have been developed which enable to switch the expression of a transgene on or off and can be combined to form sophisticated gene control networks [93]. In parallel, the understanding of how gene function relates to phenotypes is rapidly growing and has been further accelerated with the sequencing of the genomes of livestock species [94–96]. Thus, the application of transgenic technology to improve food production holds much promise for the future. Yet, transgenic technology should not be seen in isolation and as sole answer to all questions. The greatest benefit

will result from combining transgenic strategies with other effective technologies. At present, however, most consumers consider transgenic technology for food applications as risky and unsafe and thus not acceptable [97]. While this may be predominantly driven by a general unease about an unfamiliar new technology, it had profound effects on politicians, regulators, and research investors and has largely prevented its application. With the mounting pressure on food safety and security, food and lifestyle-related health problems in the context of the rapidly increasing population and a changing climate, calls are becoming louder that this promising technology can no longer be ignored [98]. To gain consumer acceptance, it will be crucial to engage into an open dialog with the public that balances perceived against real risks. Only if a robust risk benefit assessment forms the bases for a decision on the acceptability of food-producing transgenic livestock will there be a chance to realize the benefits of the technology.

Bibliography

Primary Literature

1. James C (2009) Global status of commercialized biotech/GM crops: 2009. ISAAA Brief No. 41. ISAAA, Ithaca, NY
2. Maga EA, Shoemaker CF, Rowe JD, BonDurant RH et al (2006) Production and processing of milk from transgenic goats expressing human lysozyme in the mammary gland. *J Dairy Sci* 89:518–524
3. Brundige DR, Maga EA, Klasing KC, Murray JD (2008) Lysozyme transgenic goats' milk influences gastrointestinal morphology in young pigs. *J Nutr* 138:921–926
4. Brundige DR, Maga EA, Klasing KC, Murray JD (2010) Consumption of pasteurized human lysozyme transgenic goats' milk alters serum metabolite profile in young pigs. *Transgenic Res* 19:563–574
5. van Berkel PH, Welling MM, Geerts M, van Veen HA et al (2002) Large scale production of recombinant human lactoferrin in the milk of transgenic cows. *Nat Biotechnol* 20:484–487
6. Kolb AF, Pewe L, Webster J, Perlman S et al (2001) Virus-neutralizing monoclonal antibody expressed in milk of transgenic mice provides full protection against virus-induced encephalitis. *J Virol* 75:2803–2809
7. Wall RJ, Kerr DE, Bondioli KR (1997) Transgenic dairy cattle: genetic engineering on a large scale. *J Dairy Sci* 80:2213–2224
8. Zuelke KA (1998) Transgenic modification of cows milk for value-added processing. *Reprod Fertil Dev* 10:671–676

9. Brophy B, Smolenski G, Wheeler T, Wells D et al (2003) Cloned transgenic cattle produce milk with higher levels of beta-casein and kappa-casein. *Nat Biotechnol* 21:157–162
10. Laible G, Brophy B, Knighton D, Wells DN (2007) Compositional analysis of dairy products derived from clones and cloned transgenic cattle. *Theriogenology* 67:166–177
11. Farrell HM, Malin EL, Brown EM, Qi PX (2006) Casein micelle structure: what can be learned from milk synthesis and structural biology? *Curr Opin Coll Interface Sci* 11:135–147
12. Clare DA, Swaisgood HE (2000) Bioactive milk peptides: a prospectus. *J Dairy Sci* 83:1187–1195
13. Brody EP (2000) Biological activities of bovine glycomacropeptide. *Br J Nutr* 84(suppl 1):S39–S46
14. Baranyi M, Hiripi L, Szabo L, Catunda AP et al (2007) Isolation and some effects of functional, low-phenylalanine kappa-casein expressed in the milk of transgenic rabbits. *J Biotechnol* 128:383–392
15. Sahi T (1994) Genetics and epidemiology of adult-type hypolactasia. *Scand J Gastroenterol Suppl* 202:7–20
16. Stinnakre MG, Vilotte JL, Soulier S, Mercier JC (1994) Creation and phenotypic analysis of alpha-lactalbumin-deficient mice. *Proc Natl Acad Sci USA* 91:6544–6548
17. Jost B, Vilotte JL, Duluc I, Rodeau JL et al (1999) Production of low-lactose milk by ectopic expression of intestinal lactase in the mouse mammary gland. *Nat Biotechnol* 17:160–164
18. Kay JK, Roche JR, Kolver ES, Thomson NA et al (2005) A comparison between feeding systems (pasture and TMR) and the effect of vitamin E supplementation on plasma and milk fatty acid profiles in dairy cows. *J Dairy Res* 72:322–332
19. Palmquist DL, Beaulieu AD, Barbano DM (1993) Feed and animal factors influencing milk fat composition. *J Dairy Sci* 76:1753–1771
20. Reh WA, Maga EA, Collette NM, Moyer A et al (2004) Hot topic: using a stearoyl-CoA desaturase transgene to alter milk fatty acid composition. *J Dairy Sci* 87:3510–3514
21. Simopoulos AP (1999) Essential fatty acids in health and chronic disease. *Am J Clin Nutr* 70:560S–569S
22. Kao BT, Lewis KA, DePeters EJ, Van Eenennaam AL (2006) Endogenous production and elevated levels of long-chain n-3 fatty acids in the milk of transgenic mice. *J Dairy Sci* 89:3195–3201
23. McCann JC, Ames BN (2005) Is docosahexaenoic acid, an n-3 long-chain polyunsaturated fatty acid, required for development of normal brain function? An overview of evidence from cognitive and behavioral tests in humans and animals. *Am J Clin Nutr* 82:281–295
24. Kang JX, Wang JD, Wu L, Kang ZB (2004) Transgenic mice – Fat-1 mice convert n-6 to n-3 fatty acids. *Nature* 427:504
25. Lai L, Kang JX, Li R, Wang J et al (2006) Generation of cloned transgenic pigs rich in omega-3 fatty acids. *Nat Biotechnol* 24:435–436
26. Kang JX, Leaf A (2007) Why the omega-3 piggy should go to market. *Nat Biotechnol* 25:505
27. Saeki K, Matsumoto K, Kinoshita M, Suzuki I et al (2004) Functional expression of a Delta12 fatty acid desaturase gene from spinach in transgenic pigs. *Proc Natl Acad Sci USA* 101:6361–6366
28. Hammer RE, Brinster RL, Rosenfeld MG, Evans RM et al (1985) Expression of human growth hormone-releasing factor in transgenic mice results in increased somatic growth. *Nature* 315:413–416
29. Mathews LS, Hammer RE, Behringer RR, D'Ercole AJ et al (1988) Growth enhancement of transgenic mice expressing human insulin-like growth factor I. *Endocrinology* 123:2827–2833
30. Palmiter RD, Brinster RL, Hammer RE, Trumbauer ME et al (1982) Dramatic growth of mice that develop from eggs microinjected with metallothionein-growth hormone fusion genes. *Nature* 300:611–615
31. Palmiter RD, Norstedt G, Gelinas RE, Hammer RE et al (1983) Metallothionein-human GH fusion genes stimulate growth of mice. *Science* 222:809–814
32. Chung CS, Etherton TD, Wiggins JP (1985) Stimulation of swine growth by porcine growth hormone. *J Anim Sci* 60:118–130
33. Pursel VG, Pinkert CA, Miller KF, Bolt DJ et al (1989) Genetic engineering of livestock. *Science* 244:1281–1288
34. Murray JD, Nancarrow CD, Marshall JT, Hazelton IG et al (1989) Production of transgenic merino sheep by microinjection of ovine metallothionein-ovine growth hormone fusion genes. *Reprod Fertil Dev* 1:147–155
35. Rexroad CE Jr, Hammer RE, Bolt DJ, Mayo KE et al (1989) Production of transgenic sheep with growth-regulating genes. *Mol Reprod Dev* 1:164–169
36. Pursel VG, Rexroad CE Jr (1993) Recent progress in the transgenic modification of swine and sheep. *Mol Reprod Dev* 36:251–254
37. Pursel VG, Wall RJ, Mitchell AD, Elasser TH et al (1999) Expression of insulin-like growth factor-1 in skeletal muscle of transgenic swine. In: Murray JD, Anderson GB, Oberbauer AM, McGloughlin MM (eds) *Transgenic animals in agriculture*. CABI, Oxon, UK, pp 131–144
38. Nottle MB, Nagashima H, Verma PJ, Du ZT et al (1999) Production and analysis of transgenic pigs containing a metallothionein porcine growth hormone gene construct. In: Murray JD, Anderson GB, Oberbauer AM, McGloughlin MM (eds) *Transgenic animals in agriculture*. CABI, Oxon, UK, pp 145–156
39. Adams NR, Briegel JR, Ward KA (2002) The impact of a transgene for ovine growth hormone on the performance of two breeds of sheep. *J Anim Sci* 80:2325–2333
40. Devlin RH, Yesaki YT, Biagi CA, Donaldson EM et al (1994) Extraordinary salmon growth. *Nature* 371:209–210
41. Du SJ, Gong Z, Fletcher GL, Shears MA et al (1992) Growth enhancement in transgenic Atlantic salmon by the use of an

- "all fish" chimeric growth hormone gene construct. *Biotechnology* 10:176–181
42. Nam Y, Noh J, Cho Y, Cho H et al (2001) Dramatically accelerated growth and extraordinary gigantism of transgenic mud loach *Misgurnus mizolepis*. *Transgenic Res* 10:353–362
 43. Rahman MA, Maclean N (1999) Growth performance of transgenic tilapia containing an exogenous piscine growth hormone gene. In: McAndrew B, Penman D, Hulata G (eds) *Aquaculture* 173:1/4, 333–346. 34 ref
 44. Zbikowska HM (2003) Fish can be first-advances in fish transgenesis for commercial applications. *Transgenic Res* 12:379–389
 45. Devlin RH, Biagi CA, Yesaki TY, Smailus DE et al (2001) Growth of domesticated transgenic fish. *Nature* 409:781–782
 46. McPherron AC, Lawler AM, Lee SJ (1997) Regulation of skeletal muscle mass in mice by a new TGF-beta superfamily member. *Nature* 387:83–90
 47. Kambadur R, Sharma M, Smith TP, Bass JJ (1997) Mutations in myostatin (GDF8) in double-muscled Belgian Blue and Piedmontese cattle. *Genome Res* 7:910–916
 48. McPherron AC, Lee S-J (1997) Double muscling in cattle during to mutations in the myostatin gene. *Proc Natl Acad Sci USA* 94:12457–12461
 49. Grobet L, Pirottin D, Farnir F, Poncelet D et al (2003) Modulating skeletal muscle mass by postnatal, muscle-specific inactivation of the myostatin gene. *Genesis* 35:227–238
 50. Lee S-J, Reed LA, Davies MV, Girgenrath S et al (2005) Regulation of muscle growth by multiple ligands signaling through activin type II receptors. *PNAS* 102:18117–18122
 51. Yang J, Ratovitski T, Brady J, Solomon M et al (2001) Expression of myostatin pro domain results in muscular transgenic mice. *Mol Reprod Dev* 60:351–361
 52. Jain H, Singh S, Kadam M, Sarkhel BC (2010) Knockdown of the myostatin gene by RNA interference in caprine fibroblast cells. *J Biotechnol* 145:99–102
 53. Magee TR, Artaza JN, Ferrini MG, Vernet D et al (2006) Myostatin short interfering hairpin RNA gene transfer increases skeletal muscle mass. *J Gene Med* 8:1171–1181
 54. Pirottin D, Grobet L, Adamantidis A, Farnir F et al (2005) Transgenic engineering of male-specific muscular hypertrophy. *Proc Natl Acad Sci USA* 102:6413–6418
 55. Muller M, Brem G (1998) Transgenic approaches to the increase of disease resistance in farm animals. *Revue Scientifique Et Technique (International Office Of Epizootics)* 17:365–378
 56. Whitelaw CB, Sang HM (2005) Disease-resistant genetically modified animals. *Rev Sci Tech* 24:275–283
 57. Kerr DE, Plaut K, Bramley AJ, Williamson CM et al (2001) Lysostaphin expression in mammary glands confers protection against staphylococcal infection in transgenic mice. *Nat Biotechnol* 19:66–70
 58. Wall RJ, Powell AM, Paape MJ, Kerr DE et al (2005) Genetically enhanced cows resist intramammary *Staphylococcus aureus* infection. *Nat Biotechnol* 23:445–451
 59. Seyfert HM, Henke M, Interthal H, Klussmann U et al (1996) Defining candidate genes for mastitis resistance in cattle: The role of lactoferrin and lysozyme. *J Anim Breed Genet* 113:269–276
 60. Hyvonen P, Suojala L, Orro T, Haaranen J et al (2006) Transgenic cows that produce recombinant human lactoferrin in milk are not protected from experimental *Escherichia coli* intramammary infection. *Infect Immun* 74:6206–6212
 61. Maga EA, Cullor JS, Smith W, Anderson GB et al (2006) Human lysozyme expressed in the mammary gland of transgenic dairy goats can inhibit the growth of bacteria that cause mastitis and the cold-spoilage of milk. *Foodborne Pathog Dis* 3:384–392
 62. Lo D, Pursel V, Linton PJ, Sandgren E et al (1991) Expression of Mouse IgA by transgenic mice pigs And sheep. *Eur J Immunol* 21:1001–1006
 63. Weidle UH, Lenz H, Brem G (1991) Genes encoding a mouse monoclonal antibody are expressed in transgenic mice, rabbits and pigs. *Gene* 98:185
 64. Castilla J, Pintado B, Sola I, Sanchez-Morgado JM et al (1998) Engineering passive immunity in transgenic mice secreting virus-neutralizing antibodies in milk. *Nat Biotechnol* 16:349–354
 65. Ono E, Amagai K, Taharaguchi S, Tomioka Y et al (2004) Transgenic mice expressing a soluble form of porcine nectin-1/herpesvirus entry mediator C as a model for pseudorabies-resistant livestock. *Proc Natl Acad Sci USA* 101:16150–16155
 66. Ono E, Tomioka Y, Taharaguchi S, Cherel P (2006) Comparison of protection levels against pseudorabies virus infection of transgenic mice expressing a soluble form of porcine nectin-1/HveC and vaccinated mice. *Vet Microbiol* 114:327–330
 67. Mueller M, Brenig B, Winnacker E-L, Brem G (1992) Transgenic pigs carry cDNA copies encoding the murine Mx1 protein which confers resistance to influenza virus infection. *Gene* 121:263–270
 68. Asano A, Ko JH, Morozumi T, Hamashima N et al (2002) Polymorphisms and the antiviral property of porcine Mx1 protein. *J Vet Med Sci* 64:1085–1089
 69. Baise E, Pire G, Leroy M, Gerardin J et al (2004) Conditional expression of Type I interferon-induced bovine Mx1 GTPase in a stable transgenic vero cell line interferes with replication of vesicular stomatitis virus. *J Interferon Cytokine Res* 24:513–521
 70. Leroy M, Pire G, Baise E, Desmecht D (2006) Expression of the interferon-alpha/beta-inducible bovine Mx1 dynamin interferes with replication of rabies virus. *Neurobiol Dis* 21:515
 71. Kues WA, Schwinzer R, Wirth D, Verhoeven E et al (2006) Epigenetic silencing and tissue independent expression of

- a novel tetracycline inducible system in double-transgenic pigs. *FASEB J* 20:1200–1202
72. Bueler H, Aguzzi A, Sailer A, Greiner RA et al (1993) Mice devoid of PrP are resistant to scrapie. *Cell* 73:1339–1347
 73. Perrier V, Kaneko K, Safar J, Vergara J et al (2002) Dominant-negative inhibition of prion replication in transgenic mice. *Proc Natl Acad Sci USA* 99:13079–13084
 74. Cyranoski D (2003) Koreans rustle up madness-resistant cows. *Nature* 426:743
 75. Denning C, Burl S, Ainslie A, Bracken J et al (2001) Deletion of the alpha(1, 3)galactosyl transferase (GGTA1) gene and the prion protein (PrP) gene in sheep. *Nat Biotechnol* 19:559–562
 76. Kuroiwa Y, Kasinathan P, Matsushita H, Sathiyaselan J et al (2004) Sequential targeting of the genes encoding immunoglobulin-mu and prion protein in cattle. *Nat Genet* 36:775–780
 77. Yu G, Chen J, Yu H, Liu S et al (2006) Functional disruption of the prion protein gene in cloned goats. *J Gen Virol* 87:1019–1027
 78. Golding MC, Long CR, Carmell MA, Hannon GJ et al (2006) Suppression of prion protein in livestock by RNA interference. *Proc Natl Acad Sci USA* 103:5285–5290
 79. Tobler I, Gaus SE, Deboer T, Achermann P et al (1996) Altered circadian activity rhythms and sleep in mice devoid of prion protein. *Nature* 380:639
 80. Collinge J, Whittington MA, Sidle KCL, Smith CJ et al (1994) Prion protein is necessary for normal synaptic function. *Nature* 370:295
 81. Mallucci GR, Ratte S, Asante EA, Linehan J (2002) Post-natal knockout of prion protein alters hippocampal CA1 properties, but does not result in neurodegeneration. [erratum appears in *EMBO J* 2002 Mar 1;21(5):1240]. *EMBO J* 21:202–210
 82. Criado JR, Sanchez-Alavez M, Conti B, Giacchino JL et al (2005) Mice devoid of prion protein have cognitive deficits that are rescued by reconstitution of PrP in neurons. *Neurobiol Dis* 19:255
 83. Sakaguchi S, Katamine S, Nishida N, Moriuchi R et al (1996) Loss of cerebellar Purkinje cells in aged mice homozygous for a disrupted PrP gene. *Nature* 380:528
 84. Li A, Sakaguchi S, Atarashi R, Roy BC et al (2000) Identification of a novel gene encoding a PrP-like protein expressed as chimeric transcripts fused to PrP exon 1/2 in ataxic mouse line with a disrupted PrP gene. *Cell Mol Neurobiol* 20: 553–567
 85. Moore RC, Lee IY, Silverman GL, Harrison PM et al (1999) Ataxia in prion protein (PrP)-deficient mice is associated with upregulation of the novel PrP-like protein doppel. *J Mol Biol* 292:797
 86. Rossi D, Cozzio A, Flechsig E, Klein MA et al (2001) Onset of ataxia and Purkinje cell loss in PrP null mice inversely correlated with Dpl level in brain. *EMBO J* 20:694–702
 87. Richt JA, Kasinathan P, Hamir AN, Castilla J et al (2007) Production of cattle lacking prion protein. *Nat Biotechnol* 25:132
 88. Clark J, Whitelaw B (2003) A future for transgenic livestock. *Nat Rev Genet* 4:825–833
 89. Carmona S, Ely A, Crowther C, Moolla N et al (2006) Effective inhibition of HBV replication in vivo by anti-HBx short hairpin RNAs. *Mol Ther* 13:411–421
 90. McCaffrey AP, Nakai H, Pandey K, Huang Z et al (2003) Inhibition of hepatitis B virus in mice by RNA interference. *Nat Biotechnol* 21:639–644
 91. Tompkins SM, Lo CY, Tumpey TM, Epstein SL (2004) Protection against lethal influenza virus challenge by RNA interference in vivo. *Proc Natl Acad Sci USA* 101:8682–8686
 92. Hammer RE, Pursel VG, Rexroad CE Jr, Wall RJ et al (1985) Production of transgenic rabbits, sheep and pigs by microinjection. *Nature* 315:680–683
 93. Greber D, Fussenegger M (2007) Mammalian synthetic biology: engineering of sophisticated gene networks. *J Biotechnol* 130:329–345
 94. Amaral AJ, Megens HJ, Kerstens HH, Heuven HC et al (2009) Application of massive parallel sequencing to whole genome SNP discovery in the porcine genome. *BMC Genomics* 10:374
 95. Elisk CG, Tellam RL, Worley KC, Gibbs RA et al (2009) The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* 324:522–528
 96. Wallis JW, Aerts J, Groenen MA, Crooijmans RP et al (2004) A physical map of the chicken genome. *Nature* 432:761–764
 97. Pardo R, Engelhard M, Hagen K, Jorgensen RB et al (2009) The role of means and goals in technology acceptance. A differentiated landscape of public perceptions of pharming. *EMBO Rep* 10:1069–1075
 98. Fedoroff NV, Battisti DS, Beachy RN, Cooper PJ et al (2010) Radically rethinking agriculture for the 21st century. *Science* 327:833–834

Books and Reviews

- Council for Agricultural Science and Technology (CAST) (2003) *Biotechnology in animal agriculture: an overview*. Issue Paper 23. CAST, Ames, Iowa
- Council for Agricultural Science and Technology (CAST) (2009) *Animal productivity and genetic diversity: cloned and transgenic animals*. Issue Paper 43. CAST, Ames, Iowa
- Gottlieb S, Wheeler MB (2008) *Genetically engineered animals and public health: compelling benefits for health care, nutrition, the environment and animal welfare*. www.bio.org/foodag/animals/ge_animal_benefits.pdf
- Heller KJ (2006) *Genetically engineered food: methods and detection*. Wiley-VHC, Weinheim
- Jimenez-Flores R, Richardson T (1988) Genetic engineering of the caseins to modify the behavior of milk during processing: a review. *J Dairy Sci* 71:2640–2654
- Swain DL, Charmley E, Steel JW, Coffey SG (2007) *Redesigning animal agriculture: the challenge for the 21st century*. CAB, Wallingford, UK

Transgenic Livestock, Ethical Concerns and Debate

MICKEY GJERRIS¹, REINHARD HUBER², JESPER LASSEN¹,
I ANNA S. OLSSON², PETER SANDØE¹

¹Danish Centre for Bioethics and Risk Assessment,
Institute of Food and Resource Economics, University
of Copenhagen, Frederiksberg, Denmark

²Laboratory Animal Science Group, IBMC – Instituto
de Biologia Molecular e Celular, Universidade do
Porto, Porto, Portugal

Article Outline

Glossary
Definition of the Subject
Introduction
Science and Applications
Public Perceptions of Transgenic Technologies
The Ethical Issues
Public Discussions
Future Directions
Bibliography

Glossary

Animal bioreactor Transgenic animal that produces recombinant proteins in its milk, egg white, blood, urine, or seminal plasma.

Antibody Protein produced as part of the immune reaction to render harmless a foreign substance (e.g., bacteria) entering the body of an organism.

Cloning (a) Production of exact copies (clones) of a gene/genes (gene cloning). The DNA strand containing the gene of interest is cut into suitably sized pieces (fragmentation) and the gene of interest is linked to a piece of DNA (cloning vector). This vector is then introduced into cells (transfection) which are cultured in vitro and then screened for the presence of the gene of interest. (b) Production of genetically identical organisms by somatic cell nuclear transfer (SCNT). It involves the introduction of the nucleus of a somatic cell from the organism to be cloned into an enucleated egg cell. The resulting cell divides after activation (application of electric shock) into an embryo which may then be

implanted into a surrogate mother (reproductive cloning) or used to establish a tissue culture (therapeutic cloning).

Embryonic germ cell Pluripotent stem cells derived from early germ cells with properties similar to embryonic stem cells.

Embryonic stem cell Cell derived from an early embryo that is not differentiated by itself but may divide either to form (a) other stem cells or (b) cells that differentiate into specialized cell types.

Gene copy number Genes naturally exist in varying number of copies in the genome. In relation to genetic modification, gene copy number refers to the number of copies of a transgene that integrate into the host genome.

Gene expression The assembly of a product (mainly protein) based on the information coded in a gene.

Gene targeting The modification of a certain endogenous gene of an organism based on homologous recombination.

Germ cell Cell that produces gametes (egg cells in females, sperm in males).

Heterozygous Organism/cell in which the two chromosomes in a pair contain different alleles (alternative forms of a gene) at a given locus. The dominant allele will determine the phenotype.

Homologous recombination The exchange of genetic information between two similar or identical strands of DNA (often during meiosis, i.e., the formation of gametes). This process is used for the introduction of DNA sequences into the genome of organisms by gene targeting.

Homozygous Organism/cell in which the two chromosomes in a pair contain identical alleles (alternative forms of a gene) at a given locus. The alleles may either be dominant or recessive.

Hemizygous Organism/cell with only the given allele present at the given locus of only one of the chromosomes in a pair.

Knockout The replacement of a functioning endogenous gene with an inoperable version.

Lentiviruses Viruses that are able to infect both dividing and nondividing cells and are therefore used as tools as vectors for gene delivery.

Marker-assisted selection Method allowing the selection of breeding animals based on their genotype rather than their phenotype. Regions of the genome

that control certain production traits are mapped and DNA markers that control production traits are identified. Animals whose genome contains the desired markers are selected for further breeding.

Motivation The internal state of an animal which makes it behave in a certain way; the overall summation of all internal and external factors affecting decision-making.

Pronucleus The nucleus of either a sperm (male p.) or an egg (female p.) cell before their fusion inside the egg cell in the process of fertilization. At pronuclear microinjection, the DNA containing the transgene is injected into one of the pronuclei which fuse. After implantation of the egg into a female, the resulting embryo develops into a hemizygous transgenic animal which must be bred to obtain homozygous transgenic animals.

Recombinant protein Proteins that are expressed based on recombinant DNA (rDNA). rDNA is obtained by genetic engineering techniques, i.e., the 'artificial' combination of information contained in the transgene and in the host genome.

Somatic nuclear cell transfer See Cloning.

Subjective experience Conscious mental state, an experience that the individual is aware of.

Viral vector Virus-based vehicle to introduce genetic information into cells making use of the capacity of viruses to transfer their genome into the cells they infect. For transgenesis purposes, they are rendered replication-deficient to avoid replication once the transfer of the genetic information of interest into the cell has taken place.

Definition of the Subject

Humans have been able to manipulate the genomes of livestock through selective breeding for centuries; however, direct intervention has become possible only through the development of transgenic technology over the past 3 decades. So far, genetically modified animals have mainly been developed for basic research and biomedicine, but they are slowly beginning to enter the agricultural production system. In this article, livestock is defined to be all species used within the agricultural and aquacultural system. Note that such animals can be genetically modified for use within basic and medical research as well.

Most people would readily agree that there is a difference between what humans *can* do and what they *ought* to do. Equally, most people would happily acknowledge that it is good to do the morally right thing. However, the harmony usually ends there, because although it is easy to agree that a good thing should be promoted, it is often hard to reach consensus on what that good thing is, how it can be promoted, and to what extent the end justifies the means. There may also be disagreement over what counts as a bad outcome or about whether there are acts which are wrong in themselves. As soon as questions as these are discussed whether in private or in public, people are engaging in an ethical discussion – a discussion in which the aim is to reach agreement on what is good and bad and what is right and wrong.

Such ethical discussions have closely shadowed developments within biotechnology over the past 30 years. A range of applications of biotechnology to animals, including reproductive technologies, genetic modification, and cloning, has prompted concern about the ethical limits of human use of animals. It is probably an understatement to say that discussion has so far led to no consensus in the public sphere, but it would also be an overstatement to say that the debate has led to unanimous agreement. What *has* emerged, among other things, is a clearer understanding of the basic ethical assumptions driving the different viewpoints, together with greater attention to the ethical duties of humans toward animals.

The technological challenge of animal biotechnology is: What *can* be done? This question is closely followed by the question: What *ought* to be done? As with any other human activity, this question needs to be answered if decisions on what to do are to be made. That is both the blessing and the curse of being an ethical creature equipped with a conscience. Humans cannot just act and use the possibilities that lie before them; they have to take responsibility for the values that are promoted through their choices. At the same time, there are huge disagreements about what the relationship between *can do* and *ought to do* in the area of transgenic animals is. One of the consequences of this is that ethics cannot be pursued as a mere afterthought to the technology or as something of concern only to opponents of the technology.

The subject of this article is the ethical concerns about transgenic livestock. While the strict scientific definition of “transgenesis” implies that genetic material is transferred between species [1, 2], the term is often used interchangeably with genetic modification to cover a wider range of techniques in which the animal genome is manipulated – e.g., including the inactivation of genes. For livestock applications, transgenesis proper is the predominant kind of genetic modification, and so, in this chapter, “transgenic” is used only to refer to the transfer between species. Examples of genetic modification will be used additionally to clarify important points. The goal is to make the different areas of concern that arise in the ethical debates transparent, and to discuss different ways of approaching a public dialogue on these matters.

Introduction

Gene technology and other forms of modern biotechnology have given rise to ethical controversies since they were first introduced in the early 1980s. This has led to a growing awareness among both scientists and politicians that successful implementation of biotechnology in democratic and pluralistic societies requires ethical concerns to be discussed and taken seriously. Numerous and extensive attempts have thus been made over the past 3 decades to engage stakeholders and the public in general in discussion of the ethical aspects of biotechnology.

Already, the earliest transgenic animals in the 1980s were the subject of controversy [3–6]. However, it was probably with the creation of the cloned sheep Dolly in 1997 that animal biotechnology became widely known to the public in the industrialized world [7]. Looking at both studies of public opinion on the subject and the ethical literature, one can see that the concerns here can be divided into distinct groups. Some relate to possible impacts on humans, either through health-related risks, environmental consequences, or socio-economic changes. Some revolve around the possible conflicts arising from the possibility of patenting living beings. Others revolve around animal welfare, and different concepts of that, while the last group expresses more general concerns about the (un)naturalness of the technologies, and possible violations of the integrity of the animal species. One discussion that arches over all of

these issues and has been an important determinant in framing the ethical debate as a whole focuses on how best to regulate the area.

Sociological studies have revealed public skepticism about transgenic animals. Considerations of risk, usefulness, animal welfare, and other moral values play a substantial role in the public evaluation of the technologies. Based to some extent on these studies, many attempts have been made to engage the public and stakeholders such as scientists, politicians, legislators, industry, animal welfare organizations, consumers, and others in society-wide discussions of how the new technologies can be implemented in a way that will be acceptable for a majority of citizens. Whether or not these attempts have been successful obviously depends on the criteria of success applied. What can be said, however, is that they have not led to general agreement about the development and use of transgenic animals.

In this area of ethical thinking and public discourse, it is necessary to be aware of the open-ended nature of the discussion. From a scientific point of view, there is a world of difference between a knockout pig that has had a gene disabled to better serve as a organ donor for humans (xenotransplantation) [8] and a goat that has had a human gene inserted that expresses itself in the mammary glands thus enabling human proteins to be harvested from the milk [9]. But from the point of view of the nonexpert, these differences might seem ethically unimportant if the concern in play is the possible violation of the species’ integrity.

There is also a tendency among experts to look primarily at methods, or techniques, when evaluating biotechnology, whereas members of the public often bring in considerations of usefulness as well. Ethical discussions thus differ from technical discussions, and it is necessary to critically evaluate the amount of scientific knowledge and expertise needed within different layers of the ethical discussion. Obviously, science and scientific thinking cannot be left out of ethical discussions of science and its technical applications, but neither should their importance be overstressed [10].

Similarly, it can be problematic if the subject of the ethical discussion becomes too narrow. There is a tendency in the current research environment for ethics to follow natural science into increasingly specialized fields where it becomes difficult to include other areas of science with similar concerns. But at

the level of ethical concerns, it is sometimes helpful to discuss transgenic livestock together with, for example, other issues relating to modern intensive animal production [11]. Thus the concerns that people have about transgenic livestock may to a large extent be about taking an already problematic development even further.

As a prerequisite to any ethical debate about transgenic livestock, it is of course necessary to make it clear what the subject of the discussion is, and what areas of application are involved. This ensures an informed debate in which misunderstandings resulting from unclarity are avoided.

Science and Applications

From Early Research to Maturing Technology

Humans have manipulated the genetic makeup of livestock through the process of domestication for thousands of years. Scientifically based farm animal breeding began in the early twentieth century, and, particularly since the 1950s, it has had a profound effect on most livestock species. Various modern biotechnologies – particularly reproductive technologies – have gradually been adopted since the 1980s.

A comparatively recent technology in this process of development is genetic engineering/modification; this emerged in the 1980s and has evolved considerably since. Genetic modification involves the direct manipulation of the genome of a group of founder animals, unlike traditional breeding, where animals are selected on the basis of their phenotype and then mated in order to approach a certain breeding goal gradually. There is a degree of vagueness in the definition of the techniques used to generate genetically modified organisms. Genetic modification, in the broadest sense, has been defined as an “altering of the genetic material in that [the GM] organism in a way that does not occur naturally by mating or natural recombination or both” [12]; this covers a wide array of technologies.

Genetic modification and transgenesis are sometimes used synonymously, but the term “transgenesis” is often defined more narrowly to apply to just one aspect of genetic modification: transgenic animals carry foreign genetic information (i.e., information from another species) and transmit this information to their progeny, because the transgene is integrated in

the germ cells of the animal [1]. Genetic modification may involve further modification of the genome (moving, deleting, multiplying, modifying genes) within an organism – e.g., the creation of knockout animals. Creating a knockout animal involves the targeted inactivation of a gene. In this process, the protein usually produced as a consequence of the genetic information in question is no longer produced. This is achieved by replacing an active gene with an inactive version, thus preventing the gene in question from being expressed [13]. Knockout applications are common in mice used for fundamental and biomedical research; however, in livestock species, transgenesis is the most common form of genetic modification, although even here, knockout applications do exist (e.g., mini-pigs which have had genes knocked out in connection with xenotransplantation research) [14].

The main projected advantages of transgenesis over established breeding techniques include the possibility of introducing genes that do not exist in the genome of the parents (outside the species), and the opportunity to target desirable traits without coselecting genetically coupled undesirable ones. The foreign gene is introduced in early stage embryos (first cell stage) or in cells that participate later in the complete development of the animal [13]. These embryos develop into transgenic founder animals, which can then be used in further breeding. Regardless of the particular method used, the main steps in producing transgenic animals are as follows [15]:

- Identifying the trait of interest and localizing the corresponding gene in the donor species genome
- Multiplying the gene through cloning
- Production of a suitable construct for gene transfer
- Gene transfer
- Proof of integration of the foreign gene
- Proof of expression of the foreign gene in the host animal
- Demonstrating that offspring of founder animals inherit the foreign gene
- Selective breeding to diffuse the foreign gene into the relevant population of animals

Some of the technology to create transgenic livestock is based on methods developed for use in mice: the first transgenic mammals created by pronuclear microinjection in 1980 were mice [16]. This method

was adapted in 1985 to create the first transgenic livestock (rabbits, sheep, pigs) [17].

In further developing the technologies, researchers have tried to address technical limitations, and to develop alternatives to pronuclear injection with improved efficiency (i.e., proportion of newborn animals carrying the transgene), predictability, and control of gene expression. Predictability refers to where in the host genome the transgene is integrated. When technologies in which the integration site is random are used, there may be unexpected expression patterns as a result of interference with the host genome. Technologies allowing targeted integration increase predictability by giving greater control of gene expression. Whether or not the number of copies can be controlled is also relevant, as copy number influences the level and stability of transgene expression [18].

Table 1 provides an overview of the most prominent methods of transgenesis, noting the availability for different species as well as the predictability, efficiency, and feasibility.

There are however several practical obstacles to routine production of transgenic livestock [21]. The efficiency of the methods used was very low in the

beginning, and although progress has been made here, transgenesis still cannot compete with other methods used for the genetic improvement of livestock (e.g., selective breeding in combination with artificial insemination and marker-assisted selection) on cost efficiency, practicality, and public/political acceptance. Furthermore, many traits of interest in livestock are controlled by more than one gene and are therefore difficult to improve by transgenesis. The introduction of transgenesis in livestock breeding is further complicated by the fact that, typically at any rate, very few transgenic animals are generated in the first generation. This makes diffusion into a breeding population without corresponding inbreeding depression difficult. It also means that, since the phenotypic uniformity of transgenic offspring is often limited, the evaluation of gene expression in transgenic livestock would require a large number of animals.

Fish have been popular models for research into transgenesis as they are easy to maintain/handle and highly fertile. They have easily manipulable, large eggs, which are fertilized and develop externally; fish species also usually tolerate genetic manipulation well during early development (unlike, say, rats). The most

Transgenic Livestock, Ethical Concerns and Debate. Table 1 Summary of existing transgenic approaches (Adapted from [19, 20])

Transgenic method	Availability	Integration site(s) ^a	Number of copies per site	Efficiency ^b /technical difficulty	Note
Microinjection	Most mammals ^c	Random	Variable	Low/medium	First method
Somatic cell nuclear transfer	Most mammals	Can be determined	Often single copy	Low/high	>1997
Lentivirus-based	Most mammals	Random and multiple	Single copy	High/low	Late 1990s
Sperm manipulation	Most mammals ^c	Random	Variable	Moderate/low	Successful application in few laboratories so far
ES/EG ^d cells	Mice and rats, not (yet) available in livestock	Predetermined	Often single copy	Low/NA ^e	Recently successful in Chicken (EG [91])

^aDetermination of the site of integration or random integration in the genome

^bArbitrary scale: low, <5%; moderate, >5%; high, >80%

^cAll mammals tested so far

^dEmbryonic stem/Embryonic Germ

^eNot Applicable as not used in livestock (so far)

common technique used for gene transfer is still pronuclear microinjection but, as with other livestock species, research into increased efficiency has led to the development/adaptation of new methods, including electroporation (i.e., improving the permeability of the cell membrane for DNA macromolecules by applying short electric pulses), sperm-mediated gene transfer, direct gene transfer into tissues of adult fish (via intramuscular injection of DNA or “gene gun delivery”), and the use of viral vectors [22–24].

Risks are attached to the technology, naturally, and sometimes these are unpredictable. They include the occurrence of undesirable effects of transgene expression, which should be addressed in food safety assessment schemes (like in other emerging novel foods). In particular, safety assessment should be performed regarding toxic/antinutritional effects, allergenic potential, bioactivity (e.g., hormones active *post* consumption) [25]. For a thorough discussion of different risks associated with transgenic livestock, please see “The Ethical Issues” section.

Applications of Genetic Modification to Livestock

Several explorations of potential applications of genetic modification to livestock species have been published [1, 20, 26–29], and agricultural purposes have been examined in particular [2, 30]. The authors of these studies usually distinguish between biomedical and agricultural applications. The basis of this distinction seems to be that applications in the agricultural domain facilitate the direct consumption of products from transgenic animals/their offspring by humans, while in the biomedical field, applications target human health via the development of pharmaceuticals, antibodies, transferable organs/tissues, or by studies into basic biological processes and disease mechanisms. There are products for direct consumption, such as functional foods, which however may fall into both categories [31].

The history of R&D here shows that – apart from scientific interest in basic biological questions – transgenic livestock research has undergone changes of focus and has followed funding and fashion. In this process, biomedical applications such as xenotransplantation, animals as “bioreactors” and disease models, and among agricultural applications the “Enviropig®” have emerged.

Initial developments in transgenic farm animals focused on the improvement of one of the classical production-related traits: growth [17]. The animals involved have become popularly known as “Beltsville pigs.” Ever since, there have been reports of procedures involving transgenesis for specific problems with, and questions about, livestock production (for a recent review [2]).

There have been attempts to use the technologies to improve animal health and disease resistance. There are various ways of achieving this, including passive immunization of young animals against viruses through antibody expression in milk and improvement of early growth and/or viability in piglets through increased sow milk yield when overexpressing bovine alpha-lactalbumin. Food quality and the improved usability of animal products for human consumption (e.g., the expression of human lactoferrin in milk to obtain positive effects on the immune system, reduction of lactose to circumvent intolerance, carcass and meat composition alteration toward leaner meat, and meat with higher content of omega-3 fatty acids) can also be mentioned in this context. Another production-related application aims at diminishing the “ecological footprint” of industrial pig production: this involves the expression of a bacterial phytase gene in the saliva of pigs which leads to reduced phosphorous excretion.

Moving to the broader and related area of genetic modification, gene knockout has been employed in efforts to induce disease resistance (prion diseases) [1]; it has also been used to create pigs that do not express alpha 1,3 GT in an attempt to overcome organ rejection after xenotransplantation.

Apart from its applications in biomedicine and basic biological research, transgenic technology has been used in fish to improve production parameters such as growth and food conversion, adaptation to cold environments (lowering the freezing point of body fluids through the expression of antifreeze proteins), disease resistance and prevention of the introduction of transgenes into wild populations (sterility). As with land farm animals, in 1985, the very first successful application of transgenic technology reported in fish aimed at increased growth by introducing a gene for human growth hormone (to be expressed, in this case, in goldfish). For an extensive review of applications see [24].

In a relatively recent review paper, tentative projections as to when various applications of transgenic livestock will be applied in the field are made [20]. Most of the biomedical and agricultural applications projected by Niemann and colleagues to be introduced some time beyond 2005 are still at the research stage in 2010; and, of the applications listed, only recombinant pharmaceutical proteins have actually reached the market: in 2006 the European Commission, and in 2009 the FDA in the USA, gave approval for the drug ATryn® (a recombinant form of human antithrombin, an agent that prevents blood-clotting) to be produced in the milk of transgenic goats for pharmaceutical markets. Growth-enhanced transgenic fish are often presented as the most promising candidate for the introduction of transgenic animals for food production [23, 32].

Future Perspectives

The idea of using transgenic animals in food production is controversial not only with the public, but also within the scientific community. The success of transgenic livestock for agricultural purposes has failed to meet initial expectations not least as a result of its practical limitations as compared with other breeding tools [1].

Scientists differ considerably in their assessments of the impact transgenic livestock will have on animal agriculture, and in the time-frame within which this impact will unfold. These attitudes also reflect opinions about the usefulness and feasibility of the technology; in general, biomedical applications usually meet with more optimism than agricultural applications. The following excerpts from relevant publications that can be seen in Table 2 illustrate the range of views:

Animal biotechnology is embedded in a social context and therefore affected by general trends in the societies in which it unfolds. This also means that one should be neither excessively optimistic about altruistic motives behind the technology nor excessively suspicious about possible egoistic and narrow motives. In general, one should bear in mind – when listening to all the problems that the technology might solve and how soon it will revolutionize human lives – that in western capitalist society, the worlds of research and business, science and technology, and knowledge-building and knowledge-using, have converged. Scientists are not necessarily

independent and altruistic, but human beings working in a marketplace like everybody else. They are, in other words, stakeholders. A certain amount of sound skepticism in evaluating the claims they make is therefore probably not going to be wasted [33].

It must at the same time be remembered that the potential of the technology is enormous and the range of applications almost endless – in theory, at any rate. But as the last 25–30 years have shown, it is a long way from having an idea to making it work. Sometimes, it seems that the more that is learned through science, the more it is realized how little is known, because the world in general, and molecular biology specifically, turns out to be much more complicated than initially believed [34]. When trying to assess the potential applications of transgenic livestock, there is, therefore, a large and inescapable uncertainty.

Public Perceptions of Transgenic Technologies

This section describes public perceptions of transgenic animals. As a point of departure, it should be noted that while experts, in the nature of their profession, tend to assess transgenic animals as an isolated phenomenon, the public often see these animals within a wider context. Thus most lay persons will see transgenic animals as parts of modern biotechnology. Respecting this, this section first places public perceptions of transgenic animals within a wider context of other applications of biotechnology within the agricultural and medical fields; and, second, present various aspects of the public perceptions of transgenic animals. Another aspect of the contextuality of public perceptions is the fact that perceptions should be interpreted, or understood, in the particular context in which they arise. Perceptions reflect deeper rooted cultural and/or religious values as well as passing material settings – and so they will vary with continents, countries, and religions. The following descriptions examine public attitudes in both Europe and North America, but the main focus is on Europe.

Accounts of public perceptions have two main empirical sources: qualitative interview studies or quantitative surveys. The former, which involve group interviews or individual interviews, are effective at delivering a deeper insight into the arguments and values behind specific views, but as a result of the

Transgenic Livestock, Ethical Concerns and Debate. Table 2 Examples of scientists' views on impact of transgenic livestock and relevant time frame

	Impact	Time frame
Optimistic/clear-cut predictions	"Transgenic technology allows for the stable introduction of exogenous genetic information into livestock genomes. With its ability to enhance existing or introduce entirely novel characteristics at unprecedented magnitude and speed, this emerging technology is expected to have a profound impact on the genetic improvement of livestock in the future." [2] "The overwhelming challenge to agriculture in the next century will be to feed a huge population while creating and maintaining sustainable agricultural systems. Biotechnology, including transgenic livestock, will play a key role in meeting this challenge by allowing greater precision in resource use and product design." [30]	"Sixteen years ago, transgenic animal technology was invented, and, since then, an industry has formed to exploit that technology. In the next 16 years, products created by that technology will likely be in the hands of consumers." [26]
Moderate optimism/"some time in the future"	"New methods for modifying the genome will underpin a resurgence of research using transgenic livestock. This will not only increase our understanding of basic biology in commercial species, but might also lead to the generation of animals that are more resistant to infectious disease." [1]	"We anticipate that in the near future, genetically modified animals will play a significant role in the biomedical field but that agricultural applications will develop more slowly due to the complexity of many economically important traits and to current resistance to the concept of engineered farm animals." [15] "While various transgenic concepts for the genetic improvement of livestock animals for agriculture are being evaluated the integration of this technology into practical farming systems remains some distance in the future." [2]
Critical/uncertain about time-frame	"It is not clear at all that the benefits that genetic engineering has produced hitherto compensate the potential risks of widely using these products in animal production. Putting the question in a crude way: Are we risking our health for a 5% of the benefits in milk production? Should we use transgenic pigs for a 33% less manure spread? In general, every advance of molecular genetics has been received with high expectations in the field of animal production, but the consequent application, if any, has been much more modest." [1]	"Over the past few years several commercial ventures have withdrawn from transgenic biopharming for various, usually financial, reasons. So, even though much of the groundwork has been done it is unclear what the future holds for this use [agricultural applications] of transgenic livestock. [...] It will take both a better understanding of the genomes of livestock, with the anticipated increase in candidate genes to choose from, and a major practical success before transgenic technology seriously challenges genetic improvement of livestock through selection for most conventional traits." [1] "When a procedure is challenged for some present reason, it is not uncommon to argue that in the future all these problems will be solved. [...] In the opinion of the author of this paper, the general attitude towards genetic engineering in domestic livestock should be reconsidered." [21]

limited number of participants, the interviews give a poor picture of the representation and structure of perceptions across the public at large. Quantitative surveys, by contrast, often involve 1,000 or more respondents. Their strength, as a tool, lies in the ability to produce a picture of the distribution of perceptions within a population in a country or a region.

One of the most frequently cited sources of information about public perceptions of new biotechnologies is the Eurobarometer surveys. These have been carried out regularly and simultaneously in all EU-member countries, since 1989. Since the Eurobarometers contain a number of core questions which, with small variations, have been included in all surveys, they provide real insight into the longitudinal development of public perceptions of modern biotechnology in Europe. As a supplement – but indeed not as systematically and regularly – interview studies have been carried out. These, however, are often designed to study a specific aspect of modern biotechnology, and this aspect, or topic, is seldom transgenic animals.

Perceptions of Transgenic Animals as an Application of Modern Biotechnology

A number of interesting findings emerge in the Eurobarometer surveys [35–38]. First, the surveys demonstrate that the new biotechnologies are not assessed and rejected, or accepted, en bloc; rather the public makes a balanced judgment about different applications by weighing the relevant pros and cons. One distinct result of the surveys has been the finding that if these pros and cons are categorized as matters of risk, usefulness, and moral concerns, both risks and utilities are important, but moral concerns has a veto-like character [39].

Second, the studies show that perceptions depend on the type of application. Relatively speaking, applications within the medical domain are generally accepted, whereas food-related applications are viewed with considerably more skepticism. Qualitative studies have demonstrated that this differentiation between applications is linked to the kinds of usefulness provided by the different applications [40, 41]. Thus more acceptable applications are seen as useful, in a societal sense, in virtue of their potential contribution to, for example, the alleviation of hunger, human suffering, or

environmental problems. The less acceptable applications, by contrast, are generally useful only in an economic way or to the individual.

Within this general picture, it has been found in North America as well as in Europe that applications including transgenic animals tend to be approached with skepticism [41, 42]. In the 1996 and 2002 Eurobarometers, for example, transgenic animals produced for xenotransplantation were, together with GM foods and GM crops, assessed as one of the least acceptable of seven applications [37, 43]. Despite the fact that xenotransplantation belongs to the medical area it is viewed with some skepticism by the public. One explanation of this may be another somewhat consistent finding. This is the existence of an “organism scale” – something not only indicated by the Eurobarometers, but also found in interview studies [41, 42, 44]. On this scale, acceptance is related to the distance (as it were) between the genetically manipulated organism and humans. Apes and some other mammals are located at the top of the scale, and plants and microorganisms are at the bottom. Thus the fact that the use of transgenic animals in xenotransplantation involve organisms (animals) located toward the top of the scale may contribute to the public skepticism.

Perceptions of Transgenic Animals

This section provides more details about the specific concerns present in public perceptions of transgenic animals. The presentation is structured using the now common distinction between risks, usefulness, and ethics/other moral concerns.

Transgenic Animals as a Risk In the sociology of risk, the concept of risk most often refers to any side effect of an action, practice, or technology [45]. Here, however, public perceptions of risks of transgenic animals follow the convention within studies of biotechnology, where risks are solely understood as the (unwanted) consequences for human health and the environment.

The data on transgenic animals as a risk is somewhat limited. Eurobarometer surveys include perceptions of risk related to transgenic animals in 1996 (xenotransplants and research animals) and again in 2002 (xenotransplants). In 1996 as well as in 2002,

almost 60% of the respondents to some extent agreed that the use of transgenic animals in xenotransplantation was risky [43, 46]. It is, however, impossible to determine whether this relatively high perception of risk refers to xenotransplantation as a strategy or to the transgenic animals themselves. The fact that 60% also regarded the less complex question of transgenic research animals as risky in 1996, however, indicates a considerable concern about the risks.

A qualitative study carried out in Denmark in 2002 gives an indication of the worries lying behind this risk perception [44]. First, it was found, during the exploratory interviews, that human health problems arising from the consumption of meat from transgenic or otherwise manipulated animals were not prominent issues. In discussions of GM foods in general, it transpired that there were no particular short-term concerns, mainly because people trusted the authorities and their risk assessments – but also that, by contrast, in the long run, there was some skepticism, partly because participants were not sure if the experts could foresee the long-term risks to human health. Here, references were made to the BSE crisis, which was taken to have demonstrated experts' inability to predict health consequences. In this context, some also raised concerns about the fact that the natural order of things is challenged through the use of modern biotechnologies. This natural, or god-given, order is seen as incorporating an inherent safety mechanism that is bypassed by the use of genetic technology.

Other interview studies have found similarly low levels of concern about the risks associated with the use of transgenic animals in xenotransplantation [47]. The environmental risks also seem to be considered low: essentially, it is believed that it is possible to prevent transgenic animals (unlike microorganisms and plants) from escaping and spreading their genes.

Transgenic Animals and Usefulness As noted above usefulness has several aspects: societal, individual/self-interested, and economic. Within this spectrum, the Eurobarometer surveys aim to determine the level of perceived societal usefulness by asking how useful for society the respondents find different applications. Results show that in both 1996 and 2002, about 60% of the Europeans surveyed judged xenotransplantation useful to society, and in 1996, a somewhat smaller

proportion (52%) found transgenic research animals useful. It should be stressed, however, that both applications fall within the medical area, which is known to be generally more acceptable. The interviews in Denmark confirmed this general picture, sometimes pointing to the purpose (e.g., fighting obesity) being noble, but the strategy (e.g., transgenic leaner meat from pigs) wrong, since there are well known and less controversial alternatives (e.g., obesity-curbing diets) that should be pursued first. Interview participants, however, often found themselves trapped in a dilemma between the highly valued principle that transgenic animals must serve a societal purpose, and their own interests; and sometimes this resulted in pragmatic acceptance of applications in their own interest, despite their principled rejection of the same applications.

Transgenic Animals and Other Moral Concerns In most other situations where the human use of animals is on the agenda, animal welfare is a major theme. It is striking that there is little evidence that welfare, conceived of, in the narrow sense, involving a direct effect on the well-being of the animals, is important in public perceptions of transgenic animals. In interview studies this aspect seems to be conspicuous by its absence. Instead, participants raise questions about animal integrity, or about what can be justified with regard to human treatment of animals [39, 44] – a finding that by and large is confirmed by an OECD review from 2008 [48]. In particular, the question of integrity seems to be important in the assessment of transgenic animals, and to contribute to the establishment of the ethical limits to the degree of manipulation that can be exercised over other living beings. Rather than providing final answers to the question of where such limits should be drawn, qualitative studies allow the public to voice questions such as: “Is human life worth more than that of an animal? (...) Which is more morally acceptable: waiting for a 17-year-old motorcyclist to die or using a pig as a ‘spare parts bin’” [39].

Despite such statements, in which animals are valued as such, a zoo-sociological classification [49] can, it seems, be discerned in the few studies of public perceptions of animal biotechnology. According to this classification of animals, there are different limits to what you can do to different species of animal.

The classification seems to reflect the idea that the closer an animal resembles humans, the better are the reasons (or the greater the usefulness) required to justify manipulating it. Hence, fish are not as important as cows, and cows are less important than primates.

The Ethical Issues

In this section, the ethical concerns typically raised in published discussions of animal biotechnology and in debates in the public sphere will be analyzed. The concerns will be organized into three main categories: risks to humans, risks to animals, and other moral concerns. In this way, the section follows the findings of the sociological studies and can be seen as an attempt to provide an ethical analysis of the values underlying public perceptions of biotechnology at the same time as being an assessment of the kinds of risk the development of transgenic animals present.

It is important to realize that ethical categories do not come out of the blue. The fact that many discussions of transgenic animals center on risks is no coincidence; it reflects the fact that the regulation of biotechnology is cantered on risks, as was mentioned earlier. There are differences in the way biotechnology is regulated in different parts of the world. However, the various regulatory regimes governing transgenic livestock more or less share the following requirements: (1) the livestock must not present a risk to the environment; (2) nor may it present a risk to human health; and (3) nor may it present a risk to animal health. If these three requirements are fulfilled, there is, in most regulatory regimes, no legitimate reason to set up limits to the development and use of transgenic livestock. In light of this regulatory situation, it is no wonder that discussions of transgenic livestock, like discussions regarding other forms of biotechnology, take risk issues as their starting point.

Risks to Humans

The risks to humans are uncontroversial in the sense that everybody involved in the discussion acknowledges that, to the extent that the technology presents serious risks to humans, this is a serious matter. Following the literature on the subject, it can be said that with regard to human health and the environment, transgenic technology so far does not seem to carry

significant risks to humans, whereas the socio-economic impact of the technology on such matters as food prices, agricultural production systems, and financial interests are much harder to assess.

Health Risks to humans are most often equated with the risks to human health presented by products from transgenic animals consumed as medicine or food. Only a very limited amount of research has been done in this area, since very few products have been developed. There is, however, a substantial literature on what risks should be taken into consideration when such products are evaluated. In the medical area, it is recommended that the risk assessment should follow the conventional testing of newly developed drugs. In the food area, risks arising from the modification of proteins, allergenicity, toxicology and nutritional value, and so on, will be important parameters [50].

One concern relates to the possibility that the genes transferred might cause allergic reactions in people. If a gene from a plant is transferred into an animal and that gene is responsible for the production of a protein that makes some people allergic to the plant, then that gene could very well make the same people allergic to the meat and/or milk from the animal. There are clearly ways of avoiding the most obvious problems here – by choosing genes that are not known to produce allergenic proteins, or by labeling all products from transgenic animals. But the whole area of food allergy is scientifically complicated, with many unknowns. Basically, it would seem that the risks in this area can be managed, but for each potential product careful consideration will have to be given to the question which genes should be used for modifying which animals [51].

Another issue of importance is whether the physical composition of the food product will change and thus possibly alter the nutritional value of the product. This result could be intentional or unintentional. One might attempt to genetically modify an animal to change the nutritional value of its meat and milk. It could be the amount or composition of fat in pigs, as in the case of the omega-3 pig [52] or the amount of lactoferrin in milk [53]. In such cases, it would be necessary to evaluate how these changes might affect human health in comparison with traditional products. But the changes could be unintended. Thus they might be caused by changes in genes that are not directly linked

to the nutritional value of an animal. Molecular biology is a very complex science that tells that genes are interconnected and seldom work alone. What happens to one gene can have effects elsewhere. Thus changes in genes related, for instance, to the environmental impact of an animal might have unintended side effects in that they also affect its nutritional value. Furthermore, less than perfect methods of gene transfer could also lead to unintended genetic changes. Only through effective testing and control procedures can such problems be countered in a responsible way. As no genetically modified animals have been approved for the agricultural market yet, and since only few have been developed and tested, it is very hard to say precisely how large a risk this is. It should be clear, however, that this issue will need to be addressed on a case-by-case basis. It will depend on the gene inserted and the animal modified, and it will, in all likelihood, be impossible to say anything general about the matter.

One area of biomedicine where it can be foreseen that the technology presents serious risks to humans has attracted attention. This is the xenotransplantation of pig organs. The worry is that dormant viruses in the pig genome could be transferred to humans through the organs and then “awaken” in the new environment in the human host organism, giving rise to a newly emergent disease against which humans have no natural defense. Notorious examples of emergent diseases transferred from animals to humans are the Spanish Flu, AIDS, and Avian Influenza. In the worst-case scenario just one xenotransplantation “gone wrong” could cause a pandemic [54]. The degree of scientific uncertainty here is rather high, as is generally the case in newly emerged fields of research. However, putting problems raised by the transplantation of live tissue from one species to another to one side, at present, the risks to human health presented by transgenic animals seem to be manageable.

The Environment Some of the ethical concerns connected with transgenic livestock relate particularly to the environment. One such concern is that the animals might escape and breed with wild populations, thus spreading their genes in an uncontrollable way. The most cited example here is that of transgenic fish – e.g., salmon with genetic alterations allowing faster growth. The concerns in this area can either be about

the indirect consequences this might have for humans (in this case, economic losses to the fishing industry if transgenic fish cause havoc in the wild-living species that are already under pressure from intense fishing), or direct concerns for the animals and the wider ecosystem [51, 55]. With the exception of fish, individual animals produced by biotechnological methods are rather easy to confine, and this makes it less likely that they will be a hazard for the environment. But here, as with most other concerns discussed so far, it will be necessary to evaluate the animals case by case to estimate the risks: in this case, by trying to anticipate how the animals will interact with the environment and whether this will be a threat to human interests.

Other Concerns for Risks to Human Interests

Other concerns focus on the socio-economic changes transgenic animals could bring about, especially if they are ever integrated on a large scale into farming and food production. As the technology is not yet developed, let alone introduced, in this area, it is very hard to guess what will happen. But concerns that the technologies will strengthen the movement toward large-scale farming and could deepen the divide between the developed and the developing world are not unrealistic – witness what usually happens when new technology is introduced and what has happened within the sphere of plant biotechnology. For further discussion of these concerns about the impact of biotechnology on agriculture in general [10].

Another kind of concern for humans that is sometimes raised is that the continuing reification or commodification of nature, where nature is seen exclusively as a resource to be used by humans, will harden human ethical sensitivity in general, thus ultimately causing ethical problems between humans as well. This kind of argument has been laid against the unethical use of animals since the dawn of western philosophy – as, for instance, in the work of Thomas Aquinas (c. 1225–1274). The key idea is that by being harsh to animals, humans may end up being harsh to their fellow humans. This idea is closely connected to concerns that humans, by turning what is strange and unknown into biological factories, lose an important part of what it is to be a living being that shares the world with other living beings, thus diminishing the human world [56].

In addition to these concerns, it is often suggested that if the applications of biotechnology are accepted on animals, there might be a gradual change of views on these applications and in the end their use on humans would be found acceptable. In this “slippery slope argument” the problem is not the production of transgenic animals in itself (point A), which is seen as ethically acceptable, but the production of transgenic human beings (point B), which is seen as unacceptable and claimed to be at the bottom of an inescapable slide down a moral slope. Thus point A should be avoided, although it is perfectly acceptable in itself, because it automatically leads to point B: a place that it is in no way desirable to be at.

Risks to Animal Welfare

Transgenic technologies used on livestock may cause welfare problems for the animals. As genetic modification means, by definition, the generation of animals of a genotype and phenotype not presently existing, during research and development, it is necessary to make predictions about the health and welfare of these animals, based on what is known about the gene itself and on previous experience with the technology. Genetically modified animals have so far mainly been used in biological research and as disease models. Usually, the goal of modification is to produce animals that either under- or overexpress certain genes, or which express a mutated, disease-causing human gene. In all of these cases, body function in the organism is in some way disrupted. Although the severity varies, welfare problems often accompany these disruptions.

The situation is different with transgenic livestock. Here, there is no *formal* conflict between the objective of the research and the health of the animals: ideally, animals functioning as bioreactors or xenotransplantation donors, as well as transgenic animals in food production, should be healthy. But this is not always the case, and harmful side effects may arise. For example, the “Beltsville pigs” involved in the earliest experiments suffered from extensive health and welfare problems (stress susceptibility, with only marginal improvement of the desired trait, namely growth [57]).

The problems here may arise at different stages in the development of transgenic animals: the *in vitro* technologies used at the beginning of the process may

have a negative impact on early development by disturbing early gene expression. The site of transgene integration is also relevant, as insertional mutations may lead to the loss of host gene function in the “affected” region. This risk is more frequent with those technologies where there is no control over where the transgene integrates. However, it is primarily a problem at the research stage: as such animals are not desirable as so-called “founder animals” for diffusion of the transgene into a wider population, other individuals will be selected for further breeding.

Finally, the level, and the developmental and temporal control of their expression, may mean that proteins, which are the products of transgene expression, cause problems in transgenic animal [58]. Welfare problems deriving from the activity of the transgene itself will affect much larger numbers of animals, because these problems will accompany the gene as it spreads in the population. This could be the case, for example, if animals are engineered to improve production, and if the greater production results in more prevalent production-related diseases.

Systematic welfare assessment schemes in transgenic livestock are still lacking, even though recommendations on this matter were published almost 10 years ago [58] and have been updated more recently [59]. Conclusions about the welfare of transgenic livestock have often been based, not on comprehensive studies, but rather on small-scale, or anecdotal, reports leading scientists to state, for example, that the transgenic animals “have shown no obvious abnormal phenotype” [60], are “apparently normal” [2], or “developed normally” [20].

Welfare Assessment The requirements of welfare assessment schemes in experimental design should reflect the numbers of affected animals at different stages of a transgenic program (first time generation, an increasing number of animals, and establishment of the production herds). In the early stages, with low numbers, predictions of the welfare of the animals will mostly be based on previous experience with the transgene/gene product, if available, in other species (mice); anecdotal information will be essential. As animal numbers increase quantitative research comparing groups of transgenic and control (nontransgenic) animals should be performed. With the establishment of

production herds it will not be possible to monitor the individual animal; therefore, surveillance and sampling of welfare-relevant effects of transgenesis will become more important [59].

Which aspects of the welfare of transgenic animals are assessed will depend to some extent on how animal welfare is defined. Animal welfare scientists typically base their evaluations on the clinical health and subjective experiences of animals [61]. Often, the subjective experience of animals is taken to be expressed in what animals choose. Recently, this approach has been summarized as follows: “Good welfare is defined as animals being healthy and having what they want” [62].

A broader, and complementary, perspective also includes the animal’s opportunity to engage in essential species-specific behavior [61, 63, 64]. This broader perspective is related to an even wider set of concerns about animal biotechnology wherein, the anxiety is not that the technology poses risks to animal welfare, but that it violates animal integrity and basic concepts of naturalness. This will be further discussed in section “Naturalness and Integrity”.

The Narrow and the Broad Perspective on Animal Welfare From the narrow perspective defined above, a transgenic application is problematic only if it gives rise to health problems or negative subjective experiences. From the broader perspective, the question of animal welfare is also about the extent to which the animal is allowed to fulfill its species-specific potential. The broader perspective thus points to an additional group of considerations that has to be taken into account when reflecting on animal welfare. It should be noted that the two perspectives are not mutually exclusive. Equally, they will often deliver similar recommendations.

Nevertheless, the two kinds of consideration can, on occasion, be difficult to reconcile in practice; and in that eventuality, it becomes important to clarify what kind of perspective is in play. This can be illustrated with two examples.

First, the battery cage in which laying hens are prevented from performing a range of highly motivated behaviors emerged early on as one of the most emblematic issues in the discussion of farm animal welfare. Following many years of intense debate, accompanied by research to develop and evaluate alternative

approaches, policy makers have decided to outlaw this housing system, which will be phased out in the EU by the end of 2012. However, if the problem is defined as the frustration of highly motivated behaviors, an alternative (though as yet hypothetical) solution would be to change the birds, so that they become content with the limited life in a battery cage. From a narrow perspective, there is no ethical objection to denying the animal the opportunity to behave as a bird would do in nature (as battery cage egg production does) as long as this does not negatively affect the subjective welfare of the animals, i.e., lead to negative experiences [64]. Therefore, if birds could be caused to lack motivations other than those that could be satisfied in the cage, this would be a good thing to do from the narrow perspective; it would reduce animal suffering in what is, of course, a highly profitable housing system [65, 66].

For the foreseeable future, this seems to be an academic discussion, as the technology is not, as yet, able to produce such animals – if it ever will be. To begin with, the trait to modify would have a complex genetic background, since the objective must be an animal in which one has eradicated all motivations other than those that can be satisfied in a battery cage. Second, it will be a difficult challenge to ensure that one is indeed modifying the animal into one with a restricted set of motivations rather than an animal that reacts passively, or even with apathy, to adverse conditions [67].

From the broader animal welfare perspective, on the other hand, the very idea that hens should be designed to cope well with battery cages raises serious worries and questions about the natural life of chickens, and about the experiences that constitute such a life. Thus even if it is technically possible, this kind of manipulation is highly questionable from the broader perspective.

If the battery cage case was hypothetical, the blind hen case represents a real situation. A Canadian scientist involved in poultry breeding has bred a blind egg-laying hen [68]. The blindness was thus not inflicted on living chickens, but congenital. According to the researchers, blind hens would be at less risk from feather-pecking and cannibalism, these being welfare problems typically found within free-range production systems. Again, from a narrow welfare perspective, this would seem to be a welcome solution to a rather serious welfare problem.

From the broader perspective, both the notion of deliberately breeding chickens that have such limited potential as to be content with life in a battery cage and the aim of breeding blind hens to solve production problems in the agricultural sector are seen as ethically problematic in ways that might outweigh the advantages of these ideas as perceived from the narrow perspective. Something just seems to be amiss when you deliberately create an animal with less potential than normal [69], whether or not the animal has negative experiences as a result. From this perspective, the task is not to change the animals to fit within the limits of the production system, but to adapt the production system to accommodate the needs of the animals. Implicit in this version of the broader perspective is a certain respect for the natural state of the animal.

It should be said, however, that another version of the broader perspective takes a slightly different view. While agreeing that the natural behavior of the animal is to be respected, the advocate of this view does not consider the natural behavior of the animal to be something static. Just as domesticated animals have been bred to be better adapted to housing in confinement in the past, animals today can be bred, either conventionally or through genetic modification, to be better adapted for modern-day production systems. Thus the fact that one can alter the nature of an animal by genetic modification does not constitute an ethical problem as long as one respects the nature that the animal ends up with [65, 66].

Whether animal welfare is seen from the narrow or one of the broader perspectives, the difference between traditional breeding technologies and the new biotechnological tools seems to be more of a quantitative difference in the potential of applications and associated welfare problems than a qualitative difference that creates entirely new welfare issues. The novelty or technological nature of the changes does not in itself introduce new concerns.

Other Ethical Concerns

From the concerns regarding species-specific behavior within the broader perspective, there is a connection to other concerns that are most often treated outside the discussion of animal welfare. These are concerns about the perceived unnaturalness of the technology and

possible violations of the integrity of the animals. Rather than being related to the individual animals themselves, these have to do with a more abstract and wider human perception of what an animal is. They can be said to fall within the scope of animal ethics, but outside the scope of animal welfare.

Naturalness and Integrity Technologies involving the genetic modification of animals are often labeled “unnatural” [44, 70]. As discussed previously, there might be differences in the more technical details over how unnatural the animals are in scientific terms. An animal that has had a gene inserted from another species that it would not be able to share genes with is normally seen as more unnatural than an animal which has had an extra gene from its own species inserted. But it is often argued that from a scientific perspective, this is just a matter of degree and not something requiring the label “unnatural” to be placed on products with which it is associated. There seems to be no qualitative differences between what is going on in traditional breeding and modern biotechnologies. The one follows almost logically from the other, and to label one of them “unnatural” would be to label the other likewise [71]. However, not everyone agrees with this conclusion. It has been argued that the fact that traditional breeding methods are already accepted by at least most of the public, does not necessarily mean that the next, much more advanced, step is also acceptable [72]. The growing attention to the welfare of animals produced through transgenic technologies might even serve as reason to look back on already established, and hitherto innocent, methods of animal breeding with critical eyes [34].

Another criticism of the term “unnatural” is that, although it is intuitively compelling, the normative impact of naturalness suffers from three inherent ambiguities when applied to domesticated animals. First of all, it is almost impossible to point to a stage in the development of such animals that would constitute their natural state and thus be the developmental point that should be respected [64]. Second, even without human interference, animals evolve to adapt to changing environments. Claiming that it is unnatural to change the genomes of animals can therefore, from a scientific point of view, be seen as saying that nature is unnatural. Finally, it needs to be explained why it is

assumed that what is natural is also ethically good. Cancer, earthquakes, and influenza are naturally occurring phenomena, but very few would, for that reason, deem them to be ethically good. For reasons such as these, the notion of naturalness is very often rejected as unclear and insignificant to the ethical debate [73].

These criticisms are very relevant. However, they might be seen as utterly failing to address the fundamental issues at stake in ethical debate about unnaturalness. There seem to be two different viewpoints from which the technologies can be discussed. As Anne Chapman has pointed out, the debate around unnaturalness and biotechnology tends to operate with only two very distinct concepts of nature. Nature is seen either as everything there is (leaving nothing to be unnatural) or as everything that is untouched by humans (in which case all things touched by human activity become unnatural). What is lacking is a middle ground that takes into account the fact that, from an everyday perspective, most people experience the naturalness of things as a gradually developing quality: things can be more or less natural [74]. This difference can very well be expressed in terms of the degree of control that humans exercise over natural processes. Here, it becomes obvious that although the genetic inheritance of animals is changed both through conventional breeding and through transgenic technologies, the latter, being more powerful, are more unnatural.

This understanding of the difference between natural and unnatural as being determined basically by the extent of human control can also be used as an interpretative key that helps to clarify the concerns underlying the claim that transgenic technologies violate the *integrity* of animals.

There are several definitions of the concept of animal integrity available, yet it seems that the notion still eludes clear understanding. The word is derived from the Latin “integer” meaning *wholeness, completed, untouched, unharmed*. Rutger and Heeger [75] define animal integrity as the “wholeness and intactness of the animal and its species-specific balance, as well as the capacity to sustain itself in an environment suitable to the species.” De Vries [76] interprets this definition as one mainly aiming at physical, biological, or genetic wholeness. To a large extent, violations of animal integrity will therefore also violate the welfare of the

animal [77]; and they will do so irrespective of whether animal welfare is understood from the narrow or one of the broad perspectives.

What is of special interest here, however, is whether practices that do not produce ill health, negative mental states in the animal, or prevent the animal from performing its species-specific behavior can also be seen as violation of integrity. An example, already discussed above, would be breeding animals that fits the production systems better rather than developing housing systems that satisfy the existing behavioral needs of the animals [66]. Again, one might consider changes of composition in dairy cow milk to mimic the content of human milk [78].

The debate about what integrity is, and why it is relevant, mimics the debate about naturalness in important ways. The concept of integrity is often (as in the definition proposed by Rutger and Heeger [75]) tied tightly to biological, empirical features such as behavior, the possession of an unaltered genome, and so on. The concept is then often criticized and questioned, since it makes it unclear why the integrity is especially violated through biotechnology [79].

The notion of integrity is perhaps better understood as one expressing the ethical reservations some experience when animals are commodified to an extent where the distinction between unnatural and natural breaks down, or where the power humans exercise on the animal leaves the animal as nothing more than a resource to fulfill human needs. Integrity is thus something that is experienced directly, but something also that shows up when there is a violation of it [80]. It is experienced in a phenomenological understanding of the animal – not reducing it to an object for examination or a resource to be developed, but affirming the notion of it as another “flesh-of-the-world” entity [81] sharing a “more-than-human-lifeworld” [56] with humans and all the other members of the biotic community. The claim is then that the integrity of the animal is present at this level of experience as the “completeness” of the animal before humans show up with their ability to change animals into something that can satisfy human needs. Self-evidently, an animal can be killed, but life cannot be breathen back into it. In essence, integrity is the experience that the animal is whole, complete, full, finished, when humans encounter it. Humans cannot add to it, only take away [82].

When developed along these lines, the concept of integrity loses its biological foundations and thus becomes impossible to assess from a scientific point of view. Some will probably find that the prescientific nature of the concept disqualifies it automatically; others might find that by interpreting integrity as a “zone of untouchableness” given by the sheer existence of the animal, the moral concerns about interfering in the lives of animals can be seen to make sense, even if they go beyond the animal’s welfare interests.

Respect for Autonomy As has already been mentioned, current legislation in most countries is set up to permit transgenic livestock only if this does not pose a risk to the environment and to the animals’ health; and to ensure that products originating from transgenic livestock will be allowed to enter the food chain only if they pose no risk to human health. As should be clear from the above analysis, a number of reasons why people might be skeptical about transgenic livestock can be described, and some of these will go beyond what is captured by legislation. First, some people will assess risks differently from the experts giving advice to the authorities. Second, some people will have a wider notion of what counts as a risk than that embedded in the legislation. Finally, third, some people may have moral concerns that go beyond the risk paradigm as shown earlier.

In light of this, one can ask to what extent people should have choice when they buy animal-based products. According to current legislation in the US, there is no requirement to label products derived from GM animals, whereas current EU legislation requires such products to be labeled. Here, then there will be room for a genuine ethical disagreement about whether consumers have the right to be told about the origin of animal products – and retain this right whether or not the products pose a risk. Those opposed to mandatory labeling will typically argue that such labeling may cause unfounded fears in consumers, whereas those in favor of it may argue that consumers ought to have the right to choose for themselves in line with what matters to them; and opinion polls seem to show that many consumers want to know whether or not food products are genetically modified [83].

Public Discussions

The advent of modern biotechnologies, and subsequent ethical debate, has been closely followed by a number of stakeholders: researchers, the biotech industry, politicians devising the legislative framework, NGOs within the areas of environmentalism, animal welfare, animal rights, and so forth. From the outset, it has been clear that the technologies raise ethical issues and that it is necessary to debate these. Equally, after the controversies surrounding the introduction of GM crops and products based on these crops in Europe in the 1990s, it became glaringly apparent that there was widespread skepticism about biotechnology, and that it would be necessary to create some kind of dialogue between stakeholders in the public sphere.

In this section, it is discussed how different motivations for entering this dialogue shape the dialogue itself and the subjects that are deemed relevant within it. As should be apparent from the analysis of ethical concerns provided above, it is not a shortage of subjects to talk about that characterizes the area. As in the section on public perceptions of transgenic animals, it has been necessary to widen the scope and look at the issue at a more general level to capture the trends in the discussions. There is, however, little reason to doubt that ethical discussion of the issues raised by transgenic animals will follow the pattern seen so far in adjacent technologies.

Public Participation and Dialogue

Today “dialogue with the public” is almost a mantra within the biotech community – a community that can be said to consist of all stakeholders, from the most jubilant adherents of the technologies to the most fundamentalist skeptics. The idea of public participation in the preparation of policies for emerging technologies such as biotechnology is very widely accepted. But whereas the general idea of public participation and the concept of dialogue are almost universal within the western world, the content of these notions is far from self-evident [84].

Quite what is meant by “public participation” can, generally speaking, be said to be decided by the reasons for supporting such participation in the first place. There seems to be a continuum from, on the one hand, those who support public participation because

it is a way of assuring the public of some sort of democratic or semi-democratic influence, to those who see public participation as a way of merely legitimizing the technologies in the eyes of the public. In the first case, the goal is to live up to democratic ideals of some sort without influencing the result of the participation. In the second, the point is to get the technologies accepted [85].

The influence of the goal of the public participation on the structure of the debate can most easily be seen when the concept of *dialogue* is considered. There seem to be three different understandings of the concept of dialogue in ethical discussions of biotechnology today. One is (1) a specific variety of (so to speak) *monological* dialogue. This assumes that widespread skepticism about biotechnology is based on the public's lack of knowledge and understanding of the technologies. The remedy to this is therefore knowledge transfer from the sphere of science to the public. As soon as public has been educated the skepticism will disappear.

This way of thinking, sometimes labeled “the knowledge deficit model” [86], is very widespread in the scientific community. One noteworthy example of it can be found in the report *Why clone farm animals? Goals, motives, assumptions, values and concerns among European scientists working with cloning of farm animals* [87]. Here, a group of scientists working on cloning are interviewed about the technology and various aspects of it. All say that it is very important to establish a dialogue with the public about the technology, but when directly asked what subjects would be suited for such a dialogue, they invariably answer that the public should be provided with information about the technology.

The largest problem with the knowledge deficit model is that, although it continues to thrive within the scientific community, it has been refuted in sociological studies of the public perceptions of biotechnology. What happens when citizens are provided with information about the science and possible applications is that they form an opinion about it; but there is no evidence to suggest that knowledge in general erases skepticism [44].

(2) A second form of dialogue is as monological as the first, but the exchange moves in the opposite direction – it is the public rather than the scientists now doing all the talking. Here, the task is to examine public

attitudes to biotechnological applications and figure out which will be socially acceptable. The goal is thus to ensure that the technologies pursued will not end up in the same situation as the GMOs in Europe in the 1990s. Very often, the two monological kinds of dialogue are combined, so that consumer preferences and consumer education go hand-in-hand in an attempt to produce social acceptance for the development of the technologies [85].

(3) The third concept of dialogue in the current debate sees dialogue as a way of balancing two very important considerations. One consideration is to respect the other person, whether it is a single person, a group of stakeholders, or society at large. The other is to take responsibility for one's own views of the world and attempt to do what one sincerely believes to be in the best interest of the other (whoever that may be). A classic example of this conflict can be found in the relationship between the physician and the patient. The ethical duty for the physician is to avoid taking all responsibility from the patient (paternalism as tyranny) and yet not leave the patient on his own in trying to decide upon different methods of treatment in the mistaken belief that the task of a physician is just to provide neutral information and at all costs respect the patient's right to self-determinacy (autonomy as denial of responsibility).

The concept of dialogue implies that there are two or more opinions about something, and that the people holding these opinions are willing to discuss them, holding a small window open in the back of their minds to the possibility that they may be wrong. A dialogue where it is from the outset decided that only one part of the dialogue (and that is usually the other part) could end up changing their minds is no dialogue, but could better be described as a caricature of energetic and zealous religious proselytizing.

When promoting public participation in the debates about such things as transgenic animals, it is important to realize that this is always done with a specific goal in mind – and that this goal will always influence the way the dialogue is structured. Similarly, the motivation for entering the dialogue in the first place will influence views on what questions are relevant and who should participate. These differences are also visible within different cultures where, for example, the role of the scientists and nonexperts in the

debates are understood very differently, just as the purpose of the whole exercise is understood quite differently [88].

Possible Goals of Ethical Dialogues

Motivation for entering ethical discussion of biotechnology in general and transgenic livestock in particular varies from stakeholder to stakeholder, although it is probably safe to say that all stakeholders enter the discussion because they think that the information they can bring will convince others of their viewpoints [73]. But besides the urge to win the argument, there are other and more realistic goals of ethical dialogue, and these could provide reasons for entering in the first place.

The first reason is enlightenment or clarification. Ethical debates are where humans meet and exchange their views. This gives the involved parties a chance to explain their views and listen to the explanations of others. Thus the areas of disagreement and the values that underlie them can both be clarified. This need not lead to agreement, of course; it will only do so if some of the disagreements turn out to be based on misunderstandings either of the viewpoints of the opponent or the facts of the issue. But it can lead to a better understanding of the opponent's viewpoints; and that is valuable in itself. It is valuable, first and foremost, because understanding is itself valuable, and second, because it is easier to find areas of consensus and compromise if you have an understanding of the viewpoints of others rather than just finding them irrational or blind to the ethical dimensions of the issue. Understanding why someone has certain opinions makes it easier to respect him or her, and it can create an environment around the debate that makes it easier to move on to the next step [89].

The second possible result of ethical debates is that the decisions eventually reached are socially robust. A socially robust decision is here understood as a decision that seeks to include widely held concerns about a technology while at the same time allowing it to progress in areas that are of significant benefit to society. As mentioned earlier in [section "Public Perceptions of Transgenic Technologies,"](#) it is apparent from sociological studies that, in biotechnology, there are two scales of importance. One is the scale of organism, the

other a scale of applications. The simpler the organism, the more acceptable the technologies become; and the more the application of the technology addresses issues like health, the environment, and aid to the developing world, the more acceptable it becomes [44]. Thus using transgenic technologies to produce livestock that will benefit medical research will, in all likelihood, be more socially acceptable than producing livestock that can improve productivity in agriculture.

By targeting on areas of need, developments in transgenic livestock would therefore be less problematic than they would be if the focus was on areas of economic gain. A socially robust development process will mean that certain applications are left aside in order to gain broad societal acceptance of the progress of the technologies [90].

Future Directions

The technology to produce transgenic animals is here to stay. The genie cannot be put back in the bottle. And as the technology is further developed and refined, it will, in all likelihood, become more and more economically feasible and sustainable to produce animals through these technologies. But the ethical problems will not disappear either. The issue of animal welfare as seen from a narrow perspective will continue to play a role, as will the issues around possible negative impacts on human health and the environment. Even when more data on these risks become available, it should be obvious from the discussion of GM crops that the risk evaluation will continue to be an area of controversy for many years to come.

Discussion of the socio-economic impact of the technologies will intensify. If the animals enter the agricultural sector and the food production chain, there is reason to expect that they will only strengthen the current trends within agriculture, where ever larger units with animals dominate production in a highly technological environment. This will inevitably give rise to ethical debate – but probably debate of a sort that is already taking place about how the agricultural production system should function. In the areas of naturalness and integrity, ethical disagreement will continue to flourish, especially if animals are produced for applications that are not generally regarded as important to society.

All in all, then, it is safe to say that ethical debate about transgenic livestock, and about the development of the technology, will continue to play a significant role in the future. As research and technology development becomes more and more reliant on private funding, it will be necessary to show that the development of transgenic animals will pay off investors. Finally, there is no doubt that the area will continue to be a political and ideological battle scene. It will therefore be necessary to apply methods of public participation that include many viewpoints. These methods should help to ensure socially robust development in a technology, for which the potential to create social and ethical conflicts seem as large as the technological potential.

Bibliography

- Clark J, Whitelaw B (2003) A future for transgenic livestock. *Nat Rev Genet* 4:825–833
- Laible G (2009) Enhancing livestock through genetic engineering – recent advances and future prospects. *Comp Immunol Microbiol Infect Dis* 32:123–137
- Lacy WB, Lacy LR, Busch L (1988) Agricultural biotechnology research: practices, consequences, and policy recommendations. *Agric Hum Values* 5(3):3–14
- Baile CA, Krestel-Rickert DH (1988) Will society permit the potential of genetic engineering to advance the frontiers of biology? *J Anim Sci* 66:2125–2130
- Fox MV (1989) Genetic engineering and animal welfare. *Appl Anim Behav Sci* 22(2):105–113
- Gavora JS, Lister EE (1989) Practical and ethical considerations of agricultural research assistance for the third world. *J Agric Environ Ethics* 2(4):307–322
- Wilmut I, Schnieke AE, McWhir J, Kind AJ, Campbell KH (1997) Viable offspring derived from fetal and adult mammalian cells. *Nature* 385:810–813
- Cooper DKC, Dorling A, Pierson RN III, Rees M, Seebach J, Yazer M, Ohdan H, Awwad M, Ayares D (2007) [Alpha]1, 3-galactosyltransferase gene-knockout pigs for xenotransplantation: where do we go from here? *Transplantation* 84(1):1–7
- Kling J (2009) First US approval for a transgenic animal drug. *Nat Biotechnol* 27:302–304
- Thompson PB (1997) Food biotechnology in ethical perspective. Blackie Academic & Professional, London
- Gamborg C, Gjerris M (2009) The price of responsibility. In: Gjerris M, Gamborg C, Olesen JE, Wolf J (eds) *Earth on fire. Climate change from a philosophical and ethical perspective*. University of Copenhagen, Copenhagen, Open access. Available at www.earthonfire.foi.dk
- European Commission (2001) Directive 2001/18/EC of the European Parliament and of the Council of 12 March 2001 on the deliberate release into the environment of genetically modified organisms and repealing Council Directive 90/220/EEC – Commission Declaration. European Commission, Brussels
- Houdebine LM (2009) Methods to generate transgenic animals. In: Engelhard M, Hagen K, Boysen M (eds) *Genetic engineering in livestock: new applications and interdisciplinary perspectives*. Springer-Verlag, Berlin, pp 31–48
- Lai L, Kolber-Simonds D, Park KW, Cheong HT, Greenstein JL, Im GS, Samuel M, Bonk A, Rieke A, Day BN, Murphy CN, Carter DB, Hawley RJ, Prather RS (2002) Production of α -1,3-galactosyltransferase knockout pigs by nuclear transfer cloning. *Science* 295(5557):1089–1092
- Niemann H, Kues W, Carnwath JW (2007) Transgenic farm animals: current status and perspectives for agriculture and biomedicine. In: Engelhard M, Hagen K, Boysen M (eds) *Symposium on new application of livestock genetic engineering*. Springer-Verlag, Berlin, pp 1–30
- Gordon JW, Scangos GA, Plotkin DJ, Barbosa JA, Ruddle FH (1980) Genetic transformation of mouse embryos by microinjection of purified DNA. *Proc Natl Acad Sci USA Biol Sci* 77:7380–7384
- Hammer RE, Pursel VG, Rexroad CE, Wall R, Bolt DJ, Ebert KM, Palmer RD, Brinster RL (1985) Production of transgenic rabbits, sheep and pigs by microinjection. *Nature* 315:680–683
- Kong QR, Wu ML, Huan YJ, Zhang L, Liu HY, Bou G, Luo YB, Mu YS, Liu ZH (2009) Transgene expression is associated with copy number and cytomegalovirus promoter methylation in transgenic pigs. *PLoS ONE* 4:10
- Le Provost F, Lillico S, Passet B, Young R, Whitelaw B, Vilotte JL (2009) Zinc finger nuclease technology heralds a new era in mammalian transgenesis. *Trends Biotechnol* 28:134–141
- Niemann H, Kues W, Carnwath JW (2005) Transgenic farm animals: present and future. *Rev Sci Tech* 24(1):285–298
- Blasco A (2008) The role of genetic engineering in livestock production. *Livest Sci* 113:191–201
- Foresti F (2000) Biotechnology and fish culture. *Hydrobiologia* 420:45–47
- Zbikowska HM (2003) Fish can be first – advances in fish transgenesis for commercial applications. *Transgenic Res* 12:379–389
- Rasmussen RS, Morrissey MT (2007) Biotechnology in aquaculture: transgenics and polyploidy. *Compr Rev Food Sci Food Saf* 6:2–16
- Kelly L (2005) The safety assessment of foods from transgenic and cloned animals using the comparative approach. *Rev Sci Tech* 24:61–74
- Wall RJ (1996) Transgenic livestock: progress and prospects for the future. *Theriogenology* 45(1):57–68
- Wolf E, Scherthaner W, Zakhartchenko V, Prelle K, Stojkovic M, Brem G (2000) Transgenic technology in farm animals – progress and perspectives. *Exp Physiol* 85:615–625
- Niemann H, Kues WA (2003) Application of transgenesis in livestock for agriculture and biomedicine. *Anim Reprod Sci* 79:291–317
- Wheeler MB (2007) Agricultural applications for transgenic livestock. *Trends Biotechnol* 25:204–210

30. Bremel RD, Homan EJ, Howard TH (2001) Current and future promises of transgenesis for agricultural livestock in a global marketplace. *J Dairy Sci* 84:E1–E8
31. Prather RS (2006) Cloned transgenic heart-healthy pork? *Transgenic Res* 15:405–407
32. Hallerman EM, McLean E, Fleming IA (2007) Effects of growth hormone transgenes on the behaviour and welfare of aquacultured fishes: a review identifying research needs. *Appl Anim Behav Sci* 104:265–294
33. Nightingale P, Martin P (2004) The myth of the biotech revolution. *Trends Biotechnol* 22(1):564–569
34. Gjerris M, Olsson A, Sandøe P (2006) Animal biotechnology and animal welfare. In: Council of Europe (ed) *Ethical eye – animal welfare*. Council of Europe, Strasbourg, pp 8–110
35. Durant J, Bauer J, Gaskell G (1998) *Biotechnology in the public sphere*. Science Museum, London
36. Gaskell G, Bauer M (2002) *Biotechnology 1996–2000: the years of controversy*. Science Museum, London
37. Gaskell G, Allum N, Stares S (2003) Europeans and biotechnology in 2002. *Eurobarometer* 58.0
38. Gaskell G, Allansdottir A, Allum N, Corchero C, Fischler C, Hampel J, Jackson J, Kronberger N, Mejlgard N, Revuelta G, Schreiner C, Stares S, Torgersen H, Wagner W (2006) Europeans and biotechnology in 2005: patterns and trends. *Eurobarometer* 64.3
39. Wagner W, Kronberger N, Allum N, Cheveigné de S, Diego C, Gaskell G, Heinßen M, Midden CJH, Ødegaard M, Öhman S, Rizzo B, Rusanen T, Stathopoulou A (2002) Pandora's genes – images of genes and nature. In: Bauer MW, Gaskell G (eds) *Biotechnology. The making of a global controversy*. Cambridge University Press, Cambridge, pp 244–276
40. Lassen J, Jamison A (2006) Genetic technologies meet the public: the discourses of concern. *Sci Technol Hum Values* 31(1):8–28
41. Einsiedel EF (2005) Public perceptions of transgenic animals. *Rev Sci Tech* 24(1):149–157
42. Frewer LJ, Howard C, Shepherd R (1997) Public concerns in the United Kingdom about general and specific applications of genetic engineering: risk, benefit, and ethics. *Sci Technol Hum Values* 22(1):98–124
43. Gaskell G, Bauer MW, Durant J (1998) Public perceptions of biotechnology in 1996: eurobarometer 46.1. In: Durant J, Bauer MW, Gaskell G (eds) *Biotechnology in the public sphere. A European sourcebook*. Science Museum, London, pp 189–214
44. Lassen J, Gjerris M, Sandøe P (2006) After dolly – ethical limits to the use of biotechnology on farm animals. *Theriogenology* 65:992–1004
45. Beck U (1992) *Risk society. Towards a new modernity*. Sage, London
46. European Commission (2002) *Eurobarometer 58.0*. European Commission
47. Dahl K, Sandøe P, Johnsen PF, Lassen J, Hansen AK (2003) Outline of a risk assessment: the welfare of future xeno-donor pigs. *Anim Welfare* 12:219–237
48. Rigaud N (2008) *Biotechnology: ethical and social debates*. OECD, Paris
49. Arluke A, Sanders CR (1996) *Regarding animals*. Temple University Press, Philadelphia
50. National Research Council and Institute of Medicine of the National Academies (2004) *Safety of genetically engineered foods. Approaches to assessing unintended health effects*. National Academy of Science, Washington
51. The Royal Society of Canada (2001) *Elements of precaution: recommendations for the regulation of food biotechnology in Canada. An expert panel report on the future of food biotechnology*. The Royal Society of Canada, Ottawa
52. Lai L, Jing X, Kang X, Li R, Wang J, Witt WT, Yong HY, Hao Y, Wax DM, Murphy CN, Rieke A, Samuel M, Linville ML, Korte SW, Evans RW, Starz TE, Prather RS, Dai Y (2006) Generation of cloned transgenic pigs rich in omega-3 fatty acids. *Nat Biotechnol* 24:435–436
53. Wall RJ, Kerr DE, Bondioli KR (1997) Transgenic dairy cattle: genetic engineering on a large scale. *J Dairy Sci* 80(9):2213–2224
54. Fishman JA, Patience C (2004) Xenotransplantation: infectious risks revisited. *Am J Transplant* 2004(4):1383–1390
55. PEW Initiative on Food and Biotechnology (2003) *Future fish. Issues in science and regulation of transgenic fish*. PEW Initiative on Food and Biotechnology, Washington
56. Abram D (1996) *The spell of the sensuous*. Vintage Books, New York
57. Pursel VG, Pinkert CA, Miller KF, Bolt DJ, Campbell RG, Palmiter RD, Brinster RL, Hammer RE (1989) Genetic engineering of livestock. *Science* 244:1281–1288
58. Van Reenen CG, Meuwissen THE, Hopster H, Oldenbroek K, Kruip TAM, Blokhuis HJ (2001) Transgenesis may affect farm animal welfare: a case for systematic risk assessment. *J Anim Sci* 79:1763–1779
59. Van Reenen CG (2007) Assessing the welfare of transgenic farm animals. In: Engelhard M, Hagen K, Boysen M (eds) *Symposium on new application of livestock genetic engineering*. Springer-Verlag, Berlin, pp 119–143
60. Wheeler MB (2002) Production of transgenic livestock: promise fulfilled. In: *Sixth international workshop on the biology of lactation in farm animals*, Quebec, 2002. American Society of Animal Science, pp 32–37
61. Duncan IJH, Fraser D (1997) Understanding animal welfare. In: Appleby MC, Hughes BO (eds) *Animal welfare*. CAB International, Wallingford, pp 19–31
62. Dawkins MS (2004) Using behavior to assess welfare. *Anim Welfare* 13:3–7
63. Fraser D, Weary DM, Pajor EA, Milligan BN (1997) A scientific conception of animal welfare that reflects ethical concerns. *Anim Welfare* 6:187–205
64. Appleby MC, Sandøe P (2002) Philosophical debate on the nature of well-being: implications for animal welfare. *Anim Welfare* 11(3):283–294
65. Rollin BE (1995) *Frankenstein syndrome: ethical and social issues in the genetic engineering of animals*. Cambridge University Press, Cambridge

66. Shriver A (2009) Knocking out pain in livestock: can technology succeed where morality has stalled? *Neuroethics* 2(3):115–124
67. D'Eath RB, Conington J, Lawrence AB, Olsson IAS, Sandøe P (2010) Breeding for behavioural change in farm animals: practical, economic and ethical considerations. *Anim Welfare* 19(5):17–27
68. Ali A, Cheng KM (1985) Early egg production in genetically blind (rc/rc) chickens in comparison with sighted (Rc+/rc) controls. *Poult Sci* 64(5):789–794
69. Lassen J, Sandøe P, Forkman B (2006) Happy pigs are dirty! – conflicting perspectives on animal welfare. *Livest Sci* 103(3):221–230
70. Verhoog H (2003) Naturalness and the genetic modification of animals. *Trends Biotechnol* 21(7):294–297
71. Sandøe P, Holtug N, Simonsen HB (1996) Ethical limits to domestication. *J Agric Environ Ethics* 19(5):469–493
72. Task Group on Public Perceptions of Biotechnology (1999) Ethical aspects of agricultural biotechnology. European Federation of Biotechnology, Den Haag
73. Reiss MJ, Straughan R (2000) Improving nature. The science and ethics of genetic engineering. Cambridge University Press, Cambridge
74. Chapman A (2005) Genetic engineering: the unnatural argument. *Techné* 9(2):81–93
75. Rutger B, Heeger R (1999) Inherent worth and respect for animal integrity. In: Dol M, van Vlissingen MF, Kasanmoentalib S, Visser T, Zwart H (eds) Recognizing the intrinsic value of nature. Van Gorcum, Assen, pp 41–53
76. De Vries R (2006) Genetic engineering and the integrity of animals. *J Agric Environ Ethics* 19(5):469–493
77. Vorstenbosch J (1993) The concept of integrity. Its significance for the ethical discussion on biotechnology and animals. *Livest Prod Sci* 36:109–112
78. Gottlieb S, Wheeler MD (2008) Genetically engineered animals and public health: compelling benefits for health care, nutrition, the environment, and animal welfare. BIO – The Biotechnology Organisation, Washington
79. Bovenkerk B, Brom FWA, Van Den Bergh BJ (2002) Brave new birds: the use of 'animal integrity' in animal ethics. *Hastings Cent Rep* 32:16–22
80. Cooper DE (1998) Intervention, humility and animal integrity. In: Holland A, Johnson A (eds) Animal biotechnology and ethics. Chapman & Hall, London, pp 145–155
81. Merleau-Ponty M (1969) The visible and the invisible. Northwestern University Press, Evanston
82. Løgstrup KE (1995) Vidde og prægnans. Sprogfilosofiske betragtninger. Metafysik I. 2. udg. Gyldendal, København
83. Roosen J, Lusk JL, Fox JA (2001) Consumer demand for and attitudes toward alternative beef labeling strategies in France, Germany, and the UK. *Agribusiness* 19(1):77–90
84. Nielsen AP, Lassen J, Sandøe P (2004) Involving the public – participatory methods and democratic ideals. *Glob Bioeth* 17:191–201
85. Gjerris M (2008) The three teachings of biotechnology. In: David K, Thompson P (eds) What can nanotechnology learn from biotechnology? Elsevier, Burlington, pp 91–106
86. Bauer MW, Bonfadelli H (2002) Controversy, media coverage and public knowledge. In: Bauer MW, Gaskell G (eds) Biotechnology. The making of a global controversy. Cambridge University Press, Cambridge, pp 149–175
87. Meyer G (2005) Why clone farm animals? Goals, motives, assumptions, values and concerns among European scientists working with cloning of farm animals, Project report 8. Danish Centre for Bioethics and Risk Assessment, Frederiksberg
88. Nielsen AP, Lassen J, Sandøe P (2007) Democracy at its best? The consensus conference in a cross-national perspective. *J Agric Environ Ethics* 20:13–35
89. Gjerris M (2009) This is not a hammer – on ethics and technology. In: Bedau M, Parke E (eds) Our future with protocells: the social and ethical implications of the creation of living technology. MIT Press, Boston, pp 291–306
90. Gjerris M, Sandøe P (2006) Farm animal cloning: the role of the concept of animal integrity in debating and regulating the technology. In: Kaiser M, Lien ME (eds) Ethics and the politics of food: preprints of the 6. Congress of the European Society for Agricultural and Food Ethics. Academic, Wageningen, pp 320–324
91. van de Lavoie MC, Mather-Love C, Leighton P, Diamond JH, Heyer BS, Roberts R, Zhu L, Winters-Digiaccio P, Kerchner A, Gessaro T, Swanberg S, Delany ME, Etches RJ (2006) High-grade transgenic somatic chimeras from chicken embryonic stem cells. *Mech Dev* 123:31–41

Transgenic Technologies and Increased Livestock Fertility

F. XAVIER DONADEU, CHERYL J. ASHWORTH
The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Scotland, UK

Article Outline

Glossary
Definition of the Subject and Its Importance
Introduction
The Problem of Low Fertility in Modern Livestock
The Need and Potential of Transgenic Technologies to Improve Reproductive Performance of Livestock Species
Potential Targets for Transgenic Improvement of Livestock Fertility
Future Directions
Bibliography

Glossary

Bone morphogenetic proteins (BMPs) Group of molecules belonging to the transforming growth factor family that were initially discovered by their ability to induce the formation of bone and cartilage. BMPs provide pivotal morphogenetic signals orchestrating tissue architecture throughout the body.

Calving The act of delivering a calf at the end of normal bovine pregnancy.

Energy balance In reference to animal physiology, it is the relation between energy intake from nutrient consumption and energy expenditure to maintain body functions.

Estrus The stage of the estrous cycle when a female is sexually receptive and during which ovulation occurs.

Fertility The ability to produce healthy offspring.

First-service conception rate Refers to the percentage of animals in a herd that become pregnant upon insemination on the first estrous cycle after parturition.

Genetic selection/selective breeding The process of breeding animals or plants with the goal of enhancing particular genetic traits.

Heterozygous Refers to an individual having two different alleles for a given trait.

Homozygous Refers to an individual having identical paternally and maternally inherited alleles for a given trait.

Knockin Integration of a new gene which replaces an endogenous gene.

Knockout Functional disruption of a specific gene of an organism, commonly achieved by a partial or complete deletion of the gene sequence.

Livestock Domesticated animals raised to produce commodities such as food, fiber and labor.

Mutation A change in the DNA sequence of a gene which can result in changes in the sequence and function of the encoded protein.

Ovarian follicle The structural and functional unit in the ovary within which an oocyte grows and matures until it is expelled during ovulation.

Ovulation The process by which mature ovarian follicles rupture under the influence of luteinizing hormone to release an oocyte into the oviduct.

Phenotype/phenotypic A measurable characteristic of an animal such as hair color or growth rate, and which is the product of genetics and the environment.

Point mutation A mutation involving a single nucleotide base pair in a DNA sequence.

Production trait A genetically determined characteristic of an individual of a livestock species which relates to its capacity to provide a certain product (milk, meat, fiber, eggs, work) which contributes directly to the value of the animal for the farmer.

Promoter A regulatory DNA sequence that controls the transcription of a particular gene.

Reproductive performance The productivity of an animal or herd in terms of offspring produced.

Reproductive/breeding efficiency Related to the capacity of an individual to successfully carry out all reproductive processes from gamete formation to fertilization, establishment of pregnancy and delivery of healthy offspring.

Seasonal reproduction Refers to the dependence of reproductive activity on seasonal factors, most notably changes in day length, in certain species.

Single nucleotide polymorphism DNA sequence variations involving alteration of a single nucleotide in a genome sequence.

Transgene An exogenous gene introduced into the genome of another organism.

Transgenic An animal, plant or microbe whose genetic material has been altered using an artificial process.

Zinc finger nucleases (ZFNs) Synthetic proteins consisting of an engineered DNA-binding motif characterized by multiple finger-like protrusions (finger domain) and which has been fused to the cleavage domain of the *FokI* restriction endonuclease. ZFNs can be used to induce double-stranded breaks in specific DNA sequences and thereby promote site-specific homologous recombination and targeted manipulation of genomic loci in a variety of different cell types.

Definition of the Subject and Its Importance

The possibility of using transgenic technology to promote desirable traits, including enhanced reproductive efficiency, in livestock species has attracted

considerable interest since transgenic animals were first produced. Almost 20 years ago McEvoy and associates [1] proposed that transgenic technologies could be used to increase reproductive output in food producing animals by increasing litter size in livestock or egg-laying capacity in poultry, and by reducing the limitations associated with seasonal breeding. During the intervening 2 decades, the uptake of transgenic technologies in livestock production has been limited because of technical, regulatory and ethical constraints [2]. However, the availability of novel technologies, such as Zinc finger nucleases (ZFNs) [3], which allow for greater precision and control, together with a growing need to increase livestock productivity to meet the requirements of a growing world population will likely help to realize current aspirations on use of transgenesis in animal reproduction. The application of transgenic technologies in laboratory animals has been invaluable in considerably furthering our understanding of numerous reproductive processes including gonad development and function [4–6], oogenesis and early embryogenesis [7, 8], and trophoblast gene function [9, 10], and through this increased understanding will lead to new strategies to improve reproductive efficiency of livestock species.

Introduction

Reproductive efficiency of domestic animals is critical to the sustainability of modern livestock industries. Thousands of years of domestication led to changes in the reproductive physiology of animals, for example, a dramatic reduction in the influence of season on reproduction in cattle and pigs. During modern times, intensification of farming systems coupled with the implementation of programs of genetic improvement aimed at selection for specific (non-reproductive) productivity traits had further effects on reproductive physiology often with negative consequences to breeding efficiency, with ultimately an impact on the overall productivity and sustainability of livestock industries. Meeting the increasing world demands for animal products will require increased production of healthy offspring from livestock while coping with a decrease in the availability of natural resources to maintain these animals. Achieving this within a reasonable time frame will require the use of strategies alternative, or in

addition, to traditional breeding and genetic selection. This entry will describe current limitations in reproductive performance of livestock. The potential and limitations of transgenic technologies to address such problems will then be described and specific illustrative examples will be provided.

The Problem of Low Fertility in Modern Livestock

At its 2009 World Food Summit, the United Nations Food and Agricultural Organization recognized that agricultural output will need to increase by 70% by 2050 in order to feed the world's population which is expected to exceed 9 billion in this timeframe (FAO 2009, <http://ftp.fao.org/docrep/fao/Meeting/018/k6050e.pdf>). A significant portion of that output will be derived from the demand for animal protein, especially in developing countries as they become more affluent. Considering increasing production constrains derived from climate change and competing demands on resources, the key target for the future of livestock production will be to maximize the number of offspring produced by each female animal that are also fit for purpose. This will involve three challenges, namely, to increase the numbers of offspring that a female can produce, to optimize the in utero conditions under which fetuses develop so to maximize development of appropriate phenotypes, and to bias offspring sex to reflect animal usage. Meeting these challenges will require continued advances of traditional methods and approaches, such as husbandry and genetics, as well as the broad uptake of newer and emerging biotechnological solutions such as transgenesis.

An important point to be made is that these challenges will need to be accomplished in the face of a persistent decrease in livestock fertility associated with modern farming systems in western countries, particularly affecting the dairy, pig and chicken meat industries [11–13]. In these animals, specific production traits (milk, meat) have been genetically selected at the expense of a reduction in non-selected traits such as fertility. In many cases an even more important contribution to low fertility in modern livestock herds has resulted from industry consolidation into fewer, larger production units which are managed intensively toward the primary objective of maximal economic gain.

A clear example of this is the dairy cattle industry, where a steady increase in milk productivity over the past few decades has been associated with a progressive decline in fertility to current first-service conception rates below 40%, resulting in extended calving intervals and premature culling [11, 14, 15]. According to the UK's Dairy Science Forum (2008), poor fertility in dairy herds is estimated to cost to the UK industry alone above £300 million a year, in addition to being a major welfare issue. Under current management schemes, to maximize milk production cows need to conceive soon after parturition so that the interval between lactations is minimized and a target of one-calf-a-year can be achieved. This requires timely restoration of gonadotropin secretion and ovarian activity after calving, together with adequate uterine clearance and repair, if successful follicle maturation and ovulation followed by fertilization and establishment of pregnancy is to take place by about 90–100 days [16, 17]. However, dairy cattle have been selected to efficiently partition nutrients toward milk [18] production thus creating a state of negative energy balance after parturition where reinitiation of high milk production will occur preferentially to restoration of reproductive function conducive to establishment of pregnancy [19].

Even more profound effects on reproductive activity in cattle but also in other livestock [20–22] are derived from suboptimal management of animals under intensive farming conditions including under nutrition, stress and general poor health reflected, for example, in a high incidence of lameness and mastitis, two major conditions associated with low reproductive performance of dairy cattle [11, 14]. Repeated calving itself poses an enormous risk to sustained reproductive performance because of the incidence of parturition-related or postpartum complications (e.g., dystocia, retained fetal membranes, endometritis). Poor management, together with suboptimal ovarian function during postpartum, also leads to poor estrus detection in large herds of dairy cows and other livestock leading to a dramatic decrease in the efficiency of artificial insemination-based programs [23]. Finally, several other factors can have a large impact on overall female health including subclinical infections and environmental factors such as heat stress which has direct

effects on follicle development, oocyte quality and embryo development [24, 25].

Poor management is a primary cause of low reproductive performance of livestock in developing countries. As an example, the domestic buffalo constitutes an essential farming resource to millions of families in developing countries, particularly Asia. This species has been traditionally regarded as having low reproductive efficiency, including late attainment of puberty, a seasonal pattern of reproduction, poor expression of estrus, low conception rates and long calving intervals. To a large extent, these characteristics can be attributed to environmental and managerial factors including the absence of a year-round nutrient supply, harsh environmental conditions and suboptimal human intervention [26, 27]. Different studies have shown that buffalo fertility can improve considerably by exercising proper management and feeding and that this can be enhanced by genetic selection [27, 28].

In some instances, the problem is not a lack of nutrients but over nutrition. This is the case in the broiler chicken, in which overfeeding of growth-selected animals results in excessive recruitment of ovarian follicles leading to simultaneous ovulation of multiple follicles. This alters the timing of egg laying and results in poor eggshell quality [29]. This problem has been addressed by nutrient restriction of animals which is necessary to achieve reasonable levels of egg production. This is a significant production and welfare issue as nutrient-restricted broiler breeders show clear evidence of physiological stress as well as an increased incidence of abnormal behaviors, among other problems [30].

Limitations on the reproductive performance of livestock can also be imposed by naturally evolved traits. This is the case in sheep, goats and horses, in which ovulatory activity is restricted to certain periods of the year to ensure that offspring are born at a time when food resources are plentiful [31]. Another example is the inability to selectively breed for the desired gender for a given production trait (for example females only for milk and males only for meat production) which within livestock industries leads to considerable waste and welfare concerns due to the need to cull the less useful gender.

The Need and Potential of Transgenic Technologies to Improve Reproductive Performance of Livestock Species

As outlined above, declining fertility trends in livestock can be partially reversed by improved management. This may include pharmacological control of reproductive function which, although effective in the short term, is nonetheless expensive and has limited benefits as not all animals respond adequately to hormone treatments, particularly if these are not associated with improvements in more basic aspects of management [32–34].

Interventions aimed at increasing general herd health and welfare and reducing stress can be very effective in substantially improving fertility, particularly if they include close monitoring of the reproductive status of individual animals. Optimal implementation of such measures, however, can be costly and in some cases may not be feasible and/or economically viable, particularly in developing countries.

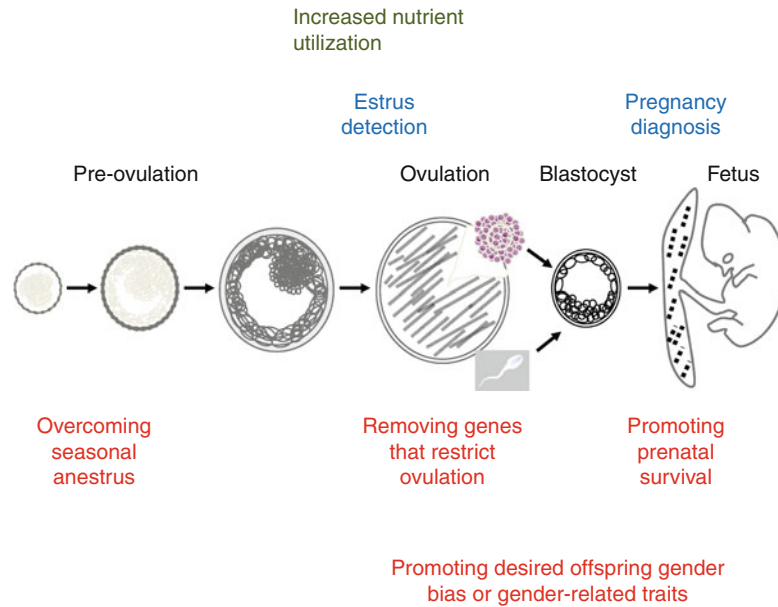
Optimum nutrition is also an important aspect of good management, particularly as it relates to prevention of low energy balance during postpartum and other critical reproductive states [15, 19]. It has been shown that the extent to which reproductive performance can be increased through nutrition is limited [35] and that a given diet can have different effects on different aspects of reproduction, as exemplified by the administration of diets which stimulate plasma insulin levels in postpartum cattle [36]. Such diets can effectively stimulate follicle development but at the same time have negative effects on the oocyte, justifying the need for carefully tailoring diet composition to specific reproductive functions if optimum fertility is to be achieved.

Although selection of livestock for certain production traits has been associated with a loss of fertility, studies with dairy cattle have shown that this loss can be reversed, through genome-wide selection on multiple traits, without completely compromising milk production [35]. However, eliminating the fertility problem in livestock through genetics would require that reproductive traits be included as a significant component of selection indexes and the desired outcome may not be achieved in decades. Alternatively, genetic gain could be aimed at reducing the impact of

poor fertility on productivity, for example, by selecting for extended lactation periods in dairy cattle [14], although in many instances this may not be economically feasible. In addition, it is unlikely that existing genetic variation in modern livestock lines will continue to generate the rate of gain obtained in the past. On that note, the use of transgenic technology to induce modifications in selected genes followed by conventional breeding and selection would offer an attractive alternative to effectively introduce desirable fertility traits in livestock, particularly in cattle and sheep which have long generation times and small offspring numbers.

Transgenic approaches have been used in livestock for the purposes of increasing productivity, conferring disease resistance and increasing environmental sustainability [2]. Similar approaches could be used to increase fertility (Fig. 1). These could be aimed at improving reproductive function directly (for example, by removing genes that naturally suppress ovulation) or indirectly (for example, by improving energy utilization by reproductive tissues) or at facilitating reproductive management of livestock (for example, by inserting reporter genes that could facilitate estrus detection or early pregnancy diagnosis). To achieve this, gene knockin or knockout approaches could be used to modify whole genes or large portions of these genes. Alternatively, novel ZFN technology allows for targeted or untargeted gene mutagenesis to be induced to selectively modify gene function in animals [37–39]. This approach would be more likely to be accepted by the public than the insertion of foreign genes into livestock genomes and it could also provide more predictable results, particularly if natural or induced mutations of the gene(s) of interest already existed in the same or a different species.

Successful application of transgenesis in livestock fertility will ultimately require a more thorough understanding of reproductive physiology in these species. Transgenic studies aimed at targeting fertility genes in livestock have not been reported and therefore potential gene targets would have to be obtained from data on naturally occurring mutations or single nucleotide polymorphisms associated with fertility, or from results of experimental gene targeting in rodents, which can already provide a substantial list of



Transgenic Technologies and Increased Livestock Fertility. Figure 1

Schematic diagram showing stages of the reproductive process where transgenic technologies could impact. The examples presented include strategies to directly improve reproductive function (in red), to indirectly favor reproductive success (in green), or to facilitate reproductive management (in blue)

candidate genes [40]. It must be stressed that, because of significant species differences, considerable caution should be used when attempting to extrapolate results from genetically modified mice into other species. A clear example of this is the failure of genetic deficiencies in bone morphogenetic protein (BMP) 15 or growth differentiation factor (GDF) 9, which are naturally associated with multiple ovulation in sheep, to increase the ovulatory response in mice [41]. In addition, it is essential to remember that most reproductive traits are controlled by several genes acting in concert and they may not therefore be easily altered by targeting a single gene. Finally, genetic modification in mice can reportedly be associated with a reduction of fertility in vitro [42], an observation that would need to be properly addressed.

The choice of a particular gene modification aimed at improving livestock fertility would need to be based on different considerations. Optimally, a gene should be targeted only in a specific cell type or organ of interest. This could be achieved with the use of cell-specific gene promoters to drive the targeting event. Choosing a gene target with naturally restricted expression, if possible

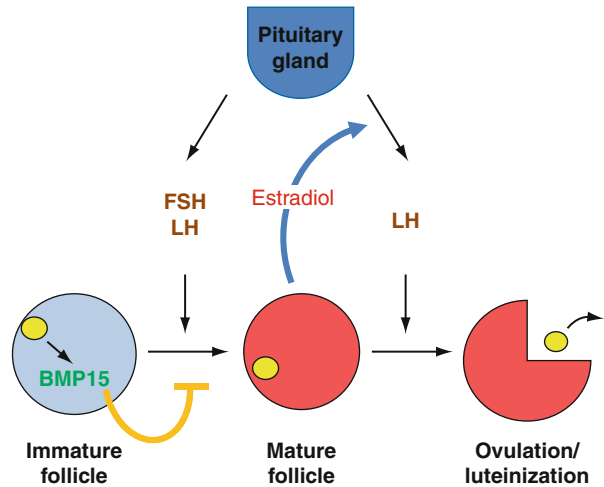
restricted to the cell type of interest, would be technically simpler and more efficient. In addition, small-scale mutagenesis rather than whole gene replacement or insertion of a new gene would be preferred to achieve the desired phenotype effects. This is now possible with remarkable efficiency in rodents and pigs by using ZFN technology [37–39]. Lastly, the intended mutation would optimally exist naturally and have been characterized thus allowing more accurate prediction of the phenotype in transgenic offspring.

Potential Targets for Transgenic Improvement of Livestock Fertility

Among the genes that could be targeted to increase fertility in livestock, certain members of the BMP family meet all three criteria described above for an optimum target. In this section, these genes will be first considered in detail followed by brief reference to others that could be targeted to improve different aspects of reproduction.

Natural mutations in the BMP ligands, BMP15 and GDF9, and in the BMP receptor type 1B (BMPRI1B)

have been described in different breeds of high prolificacy sheep and some of them have been characterized in detail [43, 44]. Five different point mutations were identified in BMP15 (*FecX* locus) and one in each of GDF9 (*FecG*) and BMPR1B (*FecB*). These mutations led in each case to reduced production of biologically active protein and/or altered protein signaling. Homozygosis for either BMP15 or GDF9 mutations is associated with infertility due to disruption of early ovarian follicle development, whereas heterozygous animals show natural higher ovulatory rates than wild types (about 3–4 vs 2 follicles per ovulation). The BMPR1B mutation in sheep is also associated with a higher ovulation rate in heterozygous animals, but in this case homozygous animals not only are fertile but they display an even larger positive effect on ovulation rate (≥ 5 follicles per ovulation). Expression of BMP15 and GDF9, but not BMPR1B, is mostly restricted to the oocyte. For both BMP15 and GDF9 mutations, the different ovarian phenotypes observed in homozygous and heterozygous animals are thought to largely reflect the role of these growth factors in promoting cellular proliferation during early stages of follicular development and in preventing premature maturation of late stage follicles, respectively (Fig. 2) [43, 45]. In heterozygotes for either mutation, reduced growth factor levels reportedly lead to increased responsiveness to the gonadotropin, follicle-stimulating hormone, and premature acquisition of responsiveness to luteinizing hormone, by granulosa cells of developing follicles which facilitates follicle maturation leading to the recruitment of an increased number of ovulatory follicles [43]. This phenomenon could be exploited using transgenesis to increase ovulatory rate not only in other sheep breeds but also in monovular species such as cattle. Evidence from immunization studies indicates that, as in sheep, altering the levels of BMP15 and/or GDF in cattle has clear effects on follicle development and ovulation rate [46]. Based on this and other information on the biological role of BMP15 and GDF9 in the bovine ovary [47], as well as on the phenotype described in mutant sheep, it can be predicted that similar mutations in BMP15 and/or GDF9 in cattle would result in animals with increased follicular sensitivity to gonadotropins and therefore a reduced gonadotropin dependence for follicle maturation. This change would provide a clear physiological



Transgenic Technologies and Increased Livestock

Fertility. Figure 2

Schematic of the physiological control of ovarian follicle maturation by pituitary gonadotropins and oocyte-derived BMP ligands. For simplicity, only BMP15 is shown. During normal estrous cycles, systemic gonadotropin levels stimulate follicle maturation leading to the production of estradiol which will eventually trigger a surge in pituitary LH release and ovulation followed by luteinization of follicular cells. Current evidence from ruminants indicates that BMP15 restricts the maturation-promoting effects of gonadotropins thus preventing the development of an excessive number of preovulatory follicles and early ovulation and/or luteinization. It is proposed that transgenically altering the levels of BMP15 in cattle would counteract the relative deficiency in gonadotropin levels during the postpartum period leading to early ovulation. *FSH* Follicle-stimulating hormone, *LH* Luteinizing hormone, *BMP15* Bone morphogenetic protein 15

advantage during situations, such as the early lactation period, when energy demands for reproduction cannot be fully met resulting in insufficient gonadotropin stimulation that prevents normal follicle maturation and restoration of ovulatory activity [16, 48]. As a result, ovulatory activity would be predictably restored early after parturition providing an opportunity for reduced calving intervals. It is very important to consider, however, that for the increased incidence of ovulation in such transgenic cattle to effectively lead to an increase in reproductive efficiency, it would have to be associated with the implementation of

measures to ensure adequate oocyte quality and uterine health, so that pregnancy could be successfully established and maintained. An increased frequency of twinning would be likely in the transgenic cattle lines which, although not desirable in certain situations, could still be exploited by the cattle industry to increase productivity. In addition to the natural mutations considered here, there are other high prolificacy phenotypes identified in sheep which do not seem to be associated with the BMP system [49]; identification of the specific genes involved will provide additional targets for manipulation of fertility in livestock.

Aside from the BMP system, there is evidence in livestock to indicate that the estrogen receptor (ESR) could be targeted to increase prolificacy. In pigs, there are associations between litter size and several genes including the ESR, prolactin receptor (PPLR), and retinol-binding protein (RBP) genes [50]. Rothschild and associates [51] found a *PvuII* polymorphism in intron 9 of *ESR1* in prolific Chinese Meishan pigs and commercial Large White populations. Among the *PvuII* genotype sows homozygous for the desired allele had larger litters at birth with more live born piglets. Introduction of a mutated or polymorphic ESR gene could therefore be used to increase litter size in several pig breeds.

The use of transgenesis has also been proposed to address gender inefficiency in livestock industries [2]. More precisely, it was suggested that genes involved in the production of functional sperm [52] or sex determination [6] and that have been successfully targeted in mice and birds could be used as targets for transgenic alteration of sex ratios in livestock. For example, the gene, *DMRT1* (doublesex and mab-3-related transcription factor 1), is implicated in male gonad development in vertebrates and its knockdown using RNA interference in early chicken embryos resulted in gonadal feminization of genetically male embryos leading to partial sex reversal [6]. An alternative approach to address gender inefficiency would be, rather than trying to alter the natural gender ratio, to transgenically enhance specific production traits in a gender-specific manner. As an example, in an attempt to produce transgenic animals with muscular hypertrophy similar to “double-muscling” which results from naturally occurring loss-of-function myostatin (*MSTN*) mutations in cattle [53], Georges and associates [54] targeted

trans-inactivators of *Mstn* on the Y chromosome of mice. The resulting transgenic males showed a 5–20% increase in skeletal muscle mass. The authors suggested the use of this approach in dairy cattle to produce breeds with combined beef and dairy abilities [54].

Finally, targeting of a number of other aspects of reproductive physiology through transgenesis may be desirable to address specific causes of reduced fertility. This could be done for example to improve prenatal survival in different species by increasing the expression of known luteotrophic and antiluteolytic conceptus signals which would be expected to reduce embryo mortality and thereby improve pregnancy rate [55], or to reduce seasonal constraints on reproduction by targeting the neuroendocrine pathways involved [56] to generate transgenic small ruminants capable of year-round breeding in temperate latitudes.

Future Directions

Since first reported over 30 years ago [57], the use of transgenic technologies in mice has been extremely useful in unraveling genetic mechanisms, including reproductive processes. Although since then the application of such technologies in other species including livestock has encountered numerous roadblocks, considerable technical improvements have been made which now allow specific and precise gene targeting in vivo [38, 39]. The vast potential of transgenic technologies in livestock has been extensively demonstrated by the successful targeting of many production-related traits in ruminants, cattle and sheep [2]. Although there has been public reluctance toward the use of such technologies in food-producing animals, technical advances allowing the production of safer foods from transgenic animals together with mounting social pressures demanding an increase in the availability of animal products for a growing world population will likely change this, and recent FDA approval of a genetically modified salmon [58] is a good indication of such change already happening. Although attempts at transgenically increasing livestock fertility have not been reported, transgenic technologies hold particular promise in that regard, both in terms of reversing current decreasing fertility trends and alleviating natural fertility constraints of livestock, two aspects that have a profound impact on overall industry

productivity and efficiency, and which is set to become increasingly important as demands for higher livestock productivity continue to increase. Studies in livestock species to increase understanding of their reproductive processes and to develop further refined gene targeting technologies will be key in bringing the realization of such promise much closer.

Bibliography

- McEvoy TG, Robinson JJ, Sreenan JM (1992) A role for transgenic animals in food production? *Trends Food Sci Technol* 3(11):294–302
- Fahrenkrug SC, Blake A, Carlson DF, Doran T, Van Eenennaam A, Faber D, Galli C, Gao Q, Hackett PB, Li N, Maga EA, Muir WM, Murray JD, Shi D, Stotish R, Sullivan E, Taylor JF, Walton M, Wheeler M, Whitelaw B, Glenn BP (2010) Precision genetics for complex objectives in animal agriculture. *J Anim Sci* 88(7):2530–2539
- Cathomen T, Keith Joung J (2008) Zinc-finger nucleases: the next generation emerges. *Mol Ther* 16(7):1200–1207
- Matzuk MM, DeMayo FJ, Hadsell LA, Kumar TR (2003) Overexpression of human chorionic gonadotropin causes multiple reproductive defects in transgenic mice. *Biol Reprod* 69(1):338–346
- Rulli SB, Ahtiainen P, Makela S, Toppari J, Poutanen M, Huhtaniemi I (2003) Elevated steroidogenesis, defective reproductive organs, and infertility in transgenic male mice overexpressing human chorionic gonadotropin. *Endocrinology* 144(11):4980–4990
- Smith CA, Roeszler KN, Ohnesorg T, Cummins DM, Farlie PG, Doran TJ, Sinclair AH (2009) The avian Z-linked gene DMRT1 is required for male sex determination in the chicken. *Nature* 461(7261):267–271
- Yan C, Elvin JA, Lin Y-N, Hadsell LA, Wang J, DeMayo FJ, Matzuk MM (2006) Regulation of growth differentiation factor 9 expression in oocytes in vivo: a key role of the E-box. *Biol Reprod* 74(6):999–1006
- Huang X, Andreu-Vieyra CV, Wang M, Cooney AJ, Matzuk MM, Zhang P (2009) Preimplantation mouse embryos depend on inhibitory phosphorylation of separase to prevent chromosome missegregation. *Mol Cell Biol* 29(6):1498–1505
- Okada Y, Ueshin Y, Isotani A, Saito-Fujita T, Nakashima H, Kimura K, Mizoguchi A, Oh-hora M, Mori Y, Ogata M, Oshima RG, Okabe M, Ikawa M (2007) Complementation of placental defects and embryonic lethality by trophoblast-specific lentiviral gene transfer. *Nat Biotechnol* 25(2):233–237
- Morioka Y, Isotani A, Oshima RG, Okabe M, Ikawa M (2009) Placenta-specific gene activation and inactivation using integrase-defective lentiviral vectors with the Cre/LoxP system. *Genesis* 47(12):793–798
- Lucy MC (2001) Reproductive loss in high-producing dairy cattle: where will it end? *J Dairy Sci* 84(6):1277–1293
- Young B, Dewey CE, Friendship RM (2010) Management factors associated with farrowing rate in commercial sow herds in Ontario. *Can Vet J* 51(2):185–189
- Julian RJ (2005) Production and growth related disorders and other metabolic diseases of poultry – a review. *Vet J* 169(3):350–369
- Dobson H, Smith RF, Royal MD, Knight CH, Sheldon IM (2007) The high producing dairy cow and its reproductive performance. *Reprod Domest Anim* 42(Suppl 2):17–23
- Chagas LM, Bass JJ, Blache D, Burke CR, Kay JK, Lindsay DR, Lucy MC, Martin GB, Meier S, Rhodes FM, Roche JR, Thatcher WW, Webb R (2007) Invited review: new perspectives on the roles of nutrition and metabolic priorities in the subfertility of high-producing dairy cows. *J Dairy Sci* 90(9):4022–4032
- Crowe M (2008) Resumption of ovarian cyclicity in postpartum beef and dairy cows. *Reprod Domest Anim* 43(s5):20–28
- Sheldon IM, Cronin J, Goetze L, Donofrio G, Schuberth H-J (2009) Defining postpartum uterine disease and the mechanisms of infection and immunity in the female reproductive tract in cattle. *Biol Reprod* 81(6):1025–1032
- Yan T, Mayne CS, Keady TWJ, Agnew RE (2006) Effects of dairy cow genotype with two planes of nutrition on energy partitioning between milk and body tissue. *J Dairy Sci* 89(3):1031–1042
- Wathes DC, Fenwick M, Cheng Z, Bourne N, Llewellyn S, Morris DG, Kenny D, Murphy J, Fitzpatrick R (2007) Influence of negative energy balance on cyclicity and fertility in the high producing dairy cow. *Theriogenology* 68:S232–S241
- von Borell E, Dobson H, Prunier A (2007) Stress, behaviour and reproductive performance in female cattle and pigs. *Horm Behav* 52(1):130–138
- Moberg GP (1991) How behavioral stress disrupts the endocrine control of reproduction in domestic animals. *J Dairy Sci* 74(1):304–311
- Reynolds LP, Borowicz PP, Caton JS, Vonnahme KA, Luther JS, Hammer CJ, Maddock Carlin KR, Grazul-Bilska AT, Redmer DA (2010) Developmental programming: the concept, large animal models, and the key role of uteroplacental vascular development. *J Anim Sci* 88(13 electronic suppl):E61–E72
- Roelofs J, López-Gatius F, Hunter RHF, van Eerdenburg FJCM, Hanzen C (2010) When is a cow in estrus? Clinical and practical aspects. *Theriogenology* 74(3):327–344
- Hansen PJ, Aréchiga CF (1999) Strategies for managing reproduction in the heat-stressed dairy cow. *J Anim Sci* 77(E-Suppl 2):36–50
- Roth Z (2008) Heat stress, the follicle, and its enclosed oocyte: mechanisms and potential strategies to improve fertility in dairy cows. *Reprod Domest Anim* 43:238–244
- Perera B (2008) Reproduction in domestic buffalo. *Reprod Domest Anim* 43:200–206
- Presicce GA (2007) Reproduction in the water buffalo. *Reprod Domest Anim* 42:24–32
- Baruselli PS, Reis EL, Marques MO, Nasser LF, Bó GA (2004) The use of hormonal treatments to improve reproductive

- performance of anestrus beef cattle in tropical climates. *Anim Reprod Sci* 82:479–486
29. Renema R, Robinson F, Proudman J, Newcombe M, McKay R (1999) Effects of body weight and feed allocation during sexual maturation in broiler breeder hens. 2. Ovarian morphology and plasma hormone profiles. *Poult Sci* 78(5):629–639
 30. Mench JA (2002) Broiler breeders: feed restriction and welfare. *World Poult Sci J* 58:23–29
 31. Donadeu FX, Watson ED (2007) Seasonal changes in ovarian activity: lessons learnt from the horse. *Anim Reprod Sci* 100 (3–4):225–242
 32. Thatcher WW, Moreira F, Santos JEP, Mattos RC, Lopes FL, Pancarci SM, Risco CA (2001) Effects of hormonal treatments on reproductive performance and embryo production. *Theriogenology* 55(1):75–89
 33. Bó GA, Baruselli PS, Chesta PM, Martins CM (2006) The timing of ovulation and insemination schedules in superstimulated cattle. *Theriogenology* 65(1):89–101
 34. De Rensis F, López-Gatius F, García-Ispuerto I, Techakumpu M (2010) Clinical use of human chorionic gonadotropin in dairy cows: an update. *Theriogenology* 73(8):1001–1008
 35. Coleman J, Pierce KM, Berry DP, Brennan A, Horan B (2009) The influence of genetic selection and feed system on the reproductive performance of spring-calving dairy cows within future pasture-based production systems. *J Dairy Sci* 92(10): 5258–5269
 36. Garnsworthy PC, Fouladi-Nashta AA, Mann GE, Sinclair KD, Webb R (2009) Effect of dietary-induced changes in plasma insulin concentrations during the early post partum period on pregnancy rate in dairy cows. *Reproduction* 137(4):759–768
 37. Cui X, Ji D, Fisher DA, Wu Y, Briner DM, Weinstein EJ (2011) Targeted integration in rat and mouse embryos with zinc-finger nucleases. *Nat Biotechnol* 29(1):64–67
 38. Whyte JJ, Zhao J, Wells KD, Samuel MS, Whitworth KM, Walters EM, Laughlin MH, Prather RS (2011) Gene targeting with zinc finger nucleases to produce cloned eGFP knockout pigs. *Mol Reprod Dev* 78(1):2
 39. Yang D, Yang H, Li W, Zhao B, Ouyang Z, Liu Z, Zhao Y, Fan N, Song J, Tian J, Li F, Zhang J, Chang L, Pei D, Chen YE, Lai L (2011) Generation of PPAR[gamma] mono-allelic knockout pigs via zinc-finger nucleases and nuclear transfer cloning. *Cell Res* 21(6):979–982
 40. Roy A, Matzuk MM (2006) Deconstructing mammalian reproduction: using knockouts to define fertility pathways. *Reproduction* 131(2):207–219
 41. Yan C, Wang P, DeMayo J, DeMayo FJ, Elvin JA, Carino C, Prasad SV, Skinner SS, Dunbar BS, Dube JL, Celeste AJ, Matzuk MM (2001) Synergistic roles of bone morphogenetic protein 15 and growth differentiation factor 9 in ovarian function. *Mol Endocrinol* 15(6):854–866
 42. Vasudevan K, Raber J, Sztein J (2010) Fertility comparison between wild type and transgenic mice by in vitro fertilization. *Transgenic Res* 19(4):587–594
 43. McNatty KP, Smith P, Moore LG, Reader K, Lun S, Hanrahan JP, Groome NP, Laitinen M, Ritvos O, Juengel JL (2005) Oocyte-expressed genes affecting ovulation rate. *Mol Cell Endocrinol* 234(1–2):57–66
 44. Souza CJH, González-Bulnes A, Campbell BK, McNeilly AS, Baird DT (2004) Mechanisms of action of the principal prolific genes and their application to sheep production. *Reprod Fertil Dev* 16(4):395–401
 45. Shimasaki S, Moore RK, Otsuka F, Erickson GF (2004) The bone morphogenetic protein system in mammalian reproduction. *Endocr Rev* 25(1):72–101
 46. Juengel JL, Hudson NL, Berg M, Hamel K, Smith P, Lawrence SB, Whiting L, McNatty KP (2009) Effects of active immunization against growth differentiation factor 9 and/or bone morphogenetic protein 15 on ovarian function in cattle. *Reproduction* 138(1):107–114
 47. Knight PG, Glistler C (2006) TGF- β superfamily members and ovarian follicle development. *Reproduction* 132(2):191–206
 48. Yavas Y, Wallon JS (2000) Induction of ovulation in postpartum suckled beef cows: a review. *Theriogenology* 54(1):1–23
 49. Juengel JL, O'Connell AR, French MC, Proctor LE, Wheeler R, Farquhar PA, Dodds KG, Galloway SM, Johnstone PD, Davis GH (2011) Identification of a line of sheep carrying a putative autosomal gene increasing ovulation rate in sheep that does not appear to interact with mutations in the TGF β superfamily. *Biol Reprod* 85(1):113–120
 50. Onteru SK, Ross JW, Rothschild MF (2009) The role of gene discovery, QTL analyses and gene expression in reproductive traits in the pig. *Soc Reprod Fertil Suppl* 66:87–102
 51. Rothschild M, Jacobson C, Vaske D, Tuggle C, Wang L, Short T, Eckardt G, Sasaki S, Vincent A, McLaren D, Southwood O, van der Steen H, Mileham A, Plastow G (1996) The estrogen receptor locus is associated with a major gene influencing litter size in pigs. *Proc Natl Acad Sci USA* 93(1):201–205
 52. Véron N, Bauer H, Weiße AY, Lüder G, Werber M, Herrmann BG (2009) Retention of gene products in syncytial spermatids promotes non-Mendelian inheritance as revealed by the t complex responder. *Genes Dev* 23(23):2705–2710
 53. Kambadur R, Sharma M, Smith TPL, Bass JJ (1997) Mutations in myostatin (GDF8) in Double-Muscling Belgian Blue and Piedmontese Cattle. *Genome Res* 7(9):910–915
 54. Pirottin D, Grobet L, Adamantidis A, Farnir F, Herens C, Schröder HD, Georges M (2005) Transgenic engineering of male-specific muscular hypertrophy. *Proc Natl Acad Sci USA* 102(18):6413–6418
 55. Bazer FW, Wu G, Spencer TE, Johnson GA, Burghardt RC, Bayless K (2010) Novel pathways for implantation and establishment and maintenance of pregnancy in mammals. *Mol Hum Reprod* 16(3):135–152
 56. Dardente H, Wyse CA, Birnie MJ, Dupré SM, Loudon ASI, Lincoln GA, Hazlerigg DG (2010) A molecular switch for photoperiod responsiveness in mammals. *Curr Biol* 20(24):2193–2198
 57. Gordon J, Ruddle F (1981) Integration and stable germ line transmission of genes injected into mouse pronuclei. *Science* 214(4526):1244–1246
 58. Fox JL (2010) Transgenic salmon inches toward finish line. *Nat Biotechnol* 28(11):1141–1142

Transgenics: Alternative Gene Transfer Methods

LOUIS-MARIE HOUEBINE

Biologie du Développement et Reproduction, Institut National de la Recherche Agronomique, Jouy en Josas, France

Article Outline

Glossary
Definition of the Subject
Introduction
Methods of Gene Transfer
Future Directions
Bibliography

Glossary

Animal cloning Non-sexual reproduction of an animal by transfer of a nucleus from a differentiated cell into the cytoplasm of an enucleated oocyte; SCNT (somatic cell nuclear transfer): cloning using a somatic cell as nuclear donor.

Chimeric animal Animal containing cells from two animals; chimeras are obtained by transferring embryonic cells from an animal into a recipient embryo; the chimera gametes contain the genome from one of the two embryos.

Hybrid Animal resulting from the sexual reproduction of parents from different breeds (intraspecies hybrid) or parents from different species (interspecies hybrid).

Knockdown Inhibition of the expression of a gene at the mRNA level by a siRNA.

Knockin Targeted integration of a gene by homologous recombination.

Knockout Inactivation of a gene by homologous recombination.

Lentiviral vector Gene construction able to integrate a foreign gene into a genome via an infection mechanism.

Meganuclease Endonuclease that is able to cut both strands of DNA and enhance the efficiency of homologous recombination.

Pluripotent cell Cell that is able to participate in the development of all the organs; lines of pluripotent

cells can be established from early embryos (ES: embryonic stem cells) or from somatic cells (iPS: induced pluripotent cells) obtained by the transfer of genes responsible for the pluripotency of embryonic cells.

siRNA Small interfering RNA (also known as RNAi) that is able to inactivate specifically an mRNA by its degradation or by the reversible inhibition of its translation; siRNAs are generated by the degradation of long double-stranded RNAs, by the transcription of micro-RNA genes, by the transcription of gene constructions coding for shRNAs (small hairpin RNAs), or by chemical synthesis.

Transgenesis Experimental transfer of an isolated gene (or a DNA fragment of any origin) to animal cells making the transmission of the genetic modification to progeny by sexual reproduction possible. The animals harboring the foreign genes are known as transgenic animals, transgenics, GM animals (genetically modified animals), rDNA animals (recombinant DNA animals), or GE animals (genetically engineered animals).

Zinc finger nuclease (ZFN) Engineered nucleases that are able to cleave both strands of genomic DNA in specific sites. ZFNs as meganucleases are able to induce a DNA repair and a local mutation equivalent to a knockout in the absence of foreign DNA or targeted gene integration at a high efficiency in the presence of a homologous recombination vector.

Definition of the Subject

Various techniques and mainly gene cloning, genome sequencing as well as transcriptomics provide researchers with an increasing number of genes. In order to know the role of the genes and the mechanisms of their regulation, it is mandatory to reintroduce them into their natural complex environment, cells, and animals. Transgenesis has thus become a major tool for biologists, and presently, at least 90% of GM animals are generated for basic studies. Transgenesis is not only a tool for research; it is also more and more extensively used for various biotechnological projects in the traditional fields of biology applications: medicine and agriculture.

The information provided by the understanding of the biological functions of humans and animals may

give clue for the design of new treatments for patients and selection of farm animals, respectively. Transgenesis has thus become a key tool for the creation of animal models for the study of human diseases. These studies may even provide researchers with genetic markers for diagnosis and genetic selection. The newly identified and studied genes may be used to produce small quantity of the corresponding recombinant proteins to establish relations between their structure and their activity. Very large quantity of recombinant proteins of pharmaceutical interest may also be produced in milk or egg white of transgenic animals. Pig cells and organs may in principle be genetically modified to become tolerated by patients. Some newly identified genes may as well be transferred to farm animal to improve production as it is already the case in plants [112](Fig. 1).

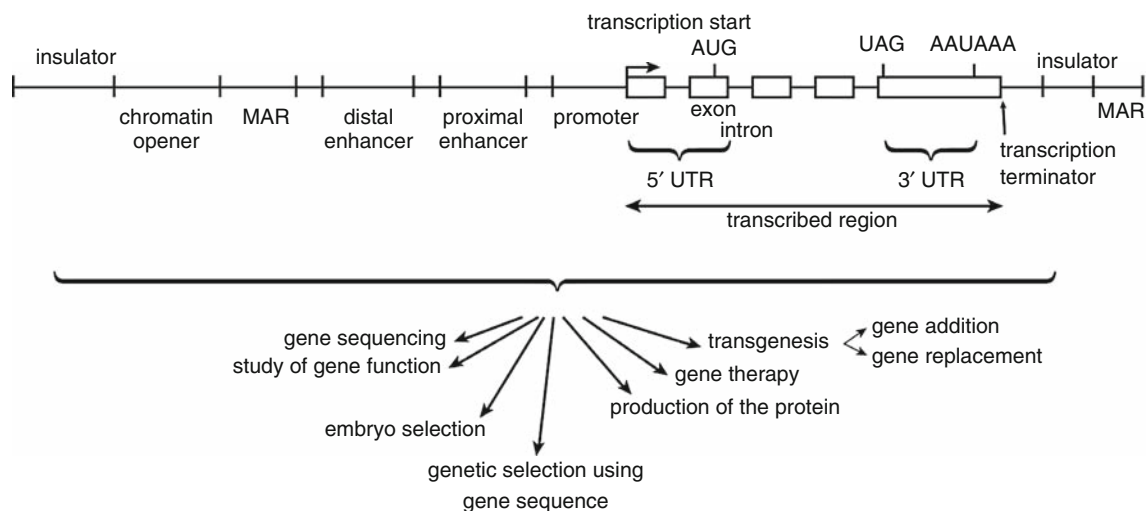
The available techniques make it possible the addition of exogenous genes and the elimination of genetic information in various animal genomes. Several more and more sophisticated tools allow also a fine-tuning of transgene expression. Although the generation of transgenic animals is no more a bottleneck for their different use, it remains laborious and costly especially in large animals. Some improvements of the methods are required and in course. The methods of gene

transfer are closely linked to the reproduction techniques and to the aim of the projects. These methods, which include the construction of the genes to be transferred and which are adapted to the different animal species are described in the present chapter.

Introduction

By essence, living organisms are in permanent evolution. This phenomenon is relatively slow, and it was probably not perceived by humans until they invented agriculture and breeding. The control of plant and animal reproduction made the empirical genetic selection possible which provided to human communities all their essential food products, pets, and ornamental plants. This led to the generation of profoundly genetically modified organisms. Carrots, tomatoes, silkworm, some dogs, etc. are unable to survive without the assistance of humans.

The discovery of the Mendel laws allowed an improvement of the genetic selection. Yet, this selection remained based on spontaneous, thus random and unknown, mutations. During the first half of the last century, it appeared necessary and possible to increase the number of random mutations to enlarge the choice of the genetically modified organisms corresponding to



Transgenics: Alternative Gene Transfer Methods. Figure 1

The different uses of isolated genes. Transgenesis, which includes random and targeted gene addition as well as specific gene inactivation and replacement is an essential tool for gene study and for biotechnological applications

the expectations of experimenters, farmers, and breeders. This was achieved by using chemical mutagens and by generating multiple intraspecies and interspecies hybrids. One of the most impressive examples is the creation of a new cereal, triticale, which results from an artificial crossing between wheat and rye. This new plant is currently a source of feed.

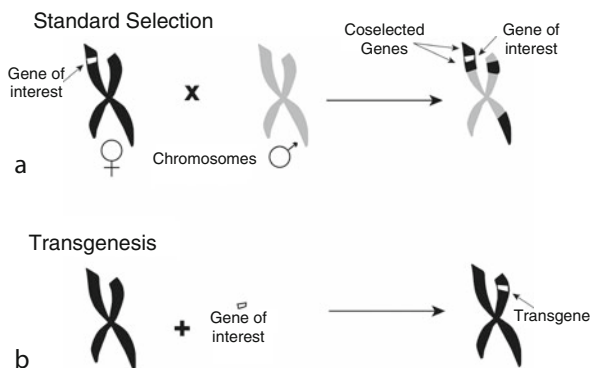
All these methods are imprecise as they induce multiple unknown mutations in addition to those which are expected. One of the problems of conventional genetic selection, either to create models or to improve production, is that the selected genes are unknown and that a large number of genes also unknown are co-selected with the genes of interest (Fig. 2). Yet, these approaches were globally highly beneficial for humans. They also show that the plasticity of living organisms is high and that humans have empirically learned to manipulate them successfully with limited undesirable side effects. The discovery of

DNA and genes opened wide avenues for research and biotechnological applications. Indeed, the manipulation of isolated and known genes makes it possible more diverse and better controlled genetic modifications (Fig. 2).

The introduction of isolated genes into cells has become a common practise in the 1970s of the last century, soon after the emergence of the genetic engineering techniques. It represented a great progress for the understanding of gene functions and mechanisms of action. This technique is still widely used, and it started being complemented in 1980 and 1983 by gene transfer into animals and plants, respectively, to generate lines of genetically modified organisms, also known as transgenic animals and plants.

The first transgenic animals, mice, were obtained by microinjecting the genes into one of the nuclei (pronuclei) of 1-day embryos [31]. This method could be extrapolated successfully to three other mammals (rabbits, pigs, and sheep) in 1985 [32], but it soon appeared that other methods had to be implemented for some species. Transgenesis is presently carried out essentially for basic research in only a few species: a mammal (mice), an insect (*Drosophila*), fish (medaka and zebra fish), and a worm (*Caenorhabditis elegans*). Some farm animals (rabbits, pigs, chicken, sheep, goat, and cow) are being used for specific studies, which cannot be performed for biological reasons with laboratory animals. Essentially, the products from some farm animals are expected to be improved by transgenesis in addition to classical genetic selection in the coming decades.

Several techniques to generate transgenic animals have made considerable progress in the past decade: (1) direct gene transfer into embryos either by microinjection of DNA, transposons, or lentiviral vectors or via sperm incubated with DNA and transferred to oocytes by ICSI (intracytoplasmic sperm injection), (2) via intermediate cells in which the gene modifications have been achieved and further used to generate animals harboring the foreign DNA sequence; these cells are either pluripotent cells that are able to give birth to chimeric transgenic animals ES (embryonic stem cells) in mice as well as iPS (induced pluripotent cells obtained after dedifferentiation of somatic cells) potentially in most species or somatic cells used for generating transgenic clones (Fig. 3).

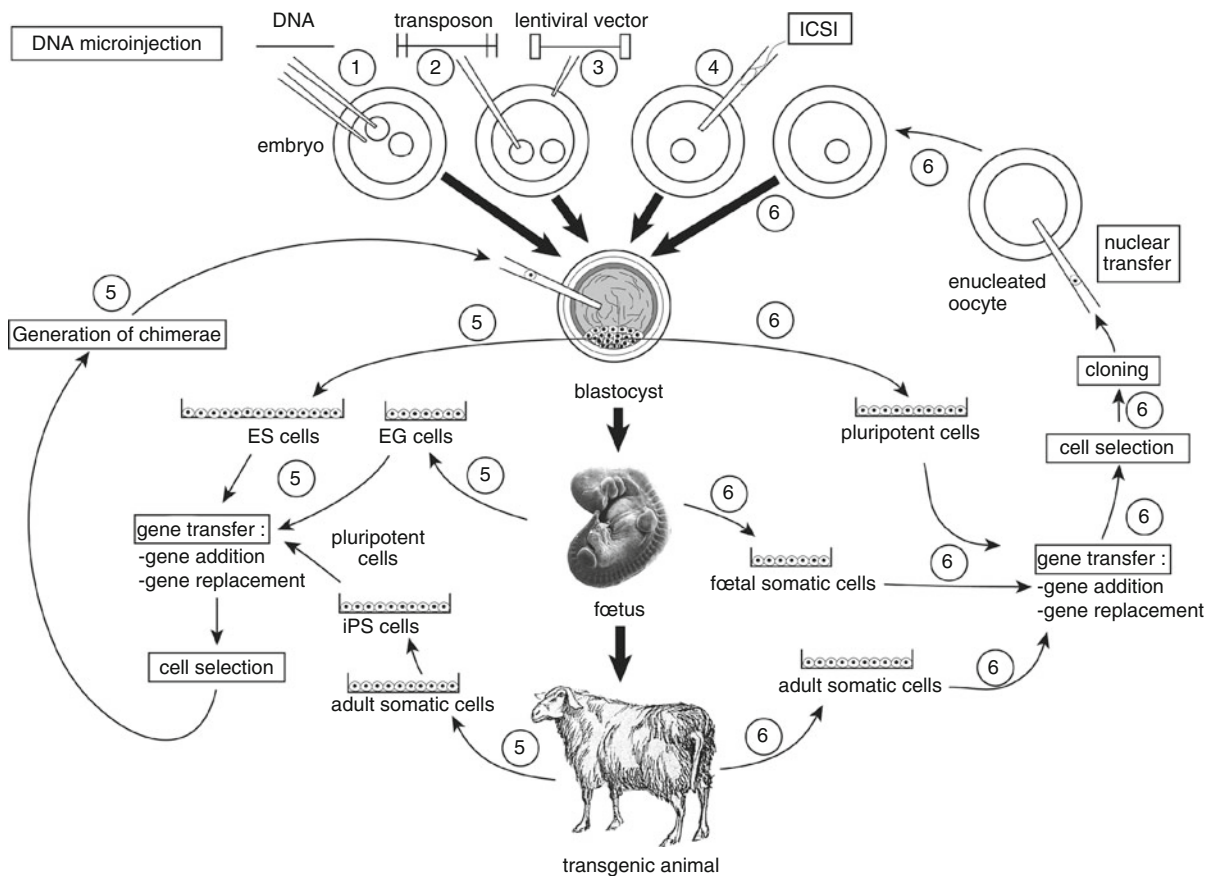


Transgenics: Alternative Gene Transfer Methods.

Figure 2

(a) Conventional genetic selection relies on the random chromosome rearrangement during sexual reproduction and, thus, to the random distribution of the different gene versions in offspring. The gene of interest responsible for the expression of a valuable genetic trait, and which is selected, is then unknown and the process implies the co-selection of a number of unknown potentially deleterious genes surrounding the gene of interest.

(b) Selection by gene transfer offers the possibility to add (or delete) a single gene having a known function into a living organism. This gene may be of various origins and may be optimized before being transferred



Transgenics: Alternative Gene Transfer Methods. Figure 3

Different methods to generate transgenic animals: (1) DNA transfer via direct microinjection into pronucleus or cytoplasm of embryo; (2) DNA transfer via a transposon: The foreign gene is introduced into the transposon, which is injected into a pronucleus; (3) DNA transfer via a lentiviral vector: The gene of interest introduced in a lentiviral vector is injected between the zona pellucida and membrane of the oocyte or the embryo; (4) DNA transfer via sperm: Sperm is incubated with the foreign gene and injected into the oocyte cytoplasm for fertilization by ICSI (intracytoplasmic sperm injection); (5) DNA transfer via pluripotent or multipotent cells: The foreign gene is introduced into pluripotent cell lines (ES, embryonic stem cells: lines established from early embryo or iPS: cells obtained after dedifferentiation of somatic cells) or into multipotent cell lines (EG, gonad cells lines established from primordial germ cells of fetal gonads); the pluripotent cells containing the foreign gene are injected into an early embryo to generate chimeric animals harboring the foreign gene DNA; the multipotent EG cells containing the foreign gene are injected into embryos to generate gametes harboring the transgene; in both cases, the transgene is transmitted to progeny; (6) DNA transfer via cloning: The foreign gene is transferred into a somatic cell, the nucleus of which is introduced into the cytoplasm of an enucleated oocyte to generate a transgenic clone. Methods 1, 2, 3, and 4 allow random gene addition whereas methods 5 and 6 allow random gene addition and targeted gene integration via homologous recombination for gene addition or gene replacement including gene knockout and knockin

Another problem emerged rapidly. The very first transgenes were expressed, and they were able, in some cases, to induce some phenotypic modifications in animals. The first example was the giant mice obtained

in 1982 [71] by overexpressing growth hormone genes. It also appeared that the expression of the transgenes was often not satisfactory and not easily controlled. This was clearly due to the insufficient knowledge of the

mechanisms controlling gene expression. Thus, for years, only empirical gene constructions having sometimes limited efficiency were used. The strategy of researchers was and often still is to generate several lines of transgenic mice (or other species) and to keep only those in which the transgene is expressed as expected. This strategy appeared insufficient when costly large transgenic animals were to be used and when finely tuned transgene expression was needed. Different methods to express transgenes in a well-controlled manner were found: (1) gene targeting by homologous recombination [8] improved or not by a genomic DNA cleavage locally induced by engineered meganucleases or zinc finger nucleases, leading either to gene knockin or knockout by a gene replacement as well as by NHEJ (non-homologous end joining), (2) conditional knockout using Cre recombinase gene controlled by exogenous inducers such as doxycycline and a engineered Cre recombinase activated by 4-hydroxy tamoxifen, (3) inactivation of mRNA by knockdown with siRNA derived from transgenes coding for shRNA or miRNA, (4) overexpression of negative transdominant proteins inactivating the genes at the protein level.

The failure of transgene expression raised fundamental questions on gene mechanism of action to experimenters. This is particularly the case for the discovery of remote regulatory elements globally known as insulators. Indeed, as opposed to bacteria, yeast, and plants, essential transcription regulators are spread over long genomic DNA regions in animals and, namely, in higher vertebrates [53]. Interestingly, answers were given to some of these questions, thanks to transgenesis and transgenesis efficiency was improved, thanks to these discoveries. Despite very significant improvement of several transgenesis methods, the efficiency of gene transfer and the control of transgene expression remain limiting factors for the optimal use of transgenic animals for research as well as for biotechnological applications. Important improvements of these methods are in course. The more efficient of gene targeting via the action of meganucleases and zinc finger endonucleases is a case in point.

Methods of Gene Transfer

Gene Transfer into Cells

Spontaneous gene transfer into cells occurs only marginally. Different techniques are being used to reach

this goal. A group of techniques relies on artificial physicochemical or mechanical processes whereas another group utilizes natural mechanisms of gene transfer, essentially infection. Transfection implies the association of DNA with molecules, which have the capacity to bind cell plasma membrane and to internalize the complex into the cytoplasm. The endocytosis process includes the transfer of the internalized material into lysosomes where it is processed or degraded. A proportion of non-degraded DNA is more or less randomly recombined. A part of the DNA is then transferred to the nucleus, and the genes it contains may be transiently transcribed. When cells do not divide, the DNA may stay several days in the nucleus. During cell division, most part of the foreign DNA is degraded and in a small percentage of cells, the foreign DNA becomes integrated into the host genome either randomly or in a targeted manner according to the sequence of the transferred DNA. A number of molecules are available, and new ones are regularly proposed to enhance the efficiency of cell transfection. These molecules may increase the endocytotic process according to the cell type, diminish the degradation of DNA, favor the transfer of DNA or the nucleus, or be less cytotoxic.

An alternative to transfection is electroporation. The method consists of incubating cells in the presence of DNA and to submit them to pulses of electric field, which generate transient pores in plasma membrane allowing the uptake of DNA. This technique is relatively efficient according to the cell type to obtain stable cells harboring a foreign gene, but many cells do not survive after the electric pulses. The fate of DNA in cells is essentially similar to this following transfection.

It is possible to select the cells in which the foreign DNA is stably integrated into the genome. To reach easily this goal, one popular way consists of co-transfecting a selection gene with the gene of interest. The selection gene may be independent of the gene of interest. Random recombination may then associate the two genes in the same DNA fragment, making the selection of the cells harboring the two co-integrated genes possible. Alternatively, the selection gene may have been associated to the gene of interest in a specific construction. The co-integration of the two genes is then more frequent, and the selection of the cells harboring the gene of interest is easier.

One possible drawback of this protocol is that the selection gene remains present in the cell.

The most efficient classical way to transfer gene into cells is the direct DNA microinjection in the cytoplasm or the nucleus. This method is relatively laborious, requires specific material, and a specific training. Moreover, gene transfer can then be achieved only in a limited number of cells.

Natural mechanisms of gene transfer are being implemented to transfer foreign DNA into cells. Some molecules are naturally able to transfer DNA from the outside of the cells to their nucleus. Lactoferrin was shown to have this property, but this mechanism seems insufficiently efficient to be used. The natural capacity of viruses to transfer their genes into cells by infection mechanisms has been retained to design viral vectors carrying the foreign genes. Such vectors have been studied and used during the last two decades with limited success for gene therapy. A few types of viruses have been retained to be used for gene therapy and those which transfer their genes into the cell genome are being implemented to generate transgenic animals. These vectors rely on the use of retroviral genomes, which must be integrated into the infected cells to be replicated and able to synthesize infectious viral particles.

Gene Transfer into Embryo

To generate transgenic animals, the foreign gene must be present and integrated into the genome of the embryos at the one-cell stage to allow its transmission into all the cells of the animal or at least of their gametes. The simplest theoretical way to reach this goal is to transfer the foreign gene directly into the embryo. In practice, this strategy is not always efficient. Indeed, embryos are rare and costly cells especially in large farm animals. In species in which gametes are very abundant, the direct gene transfer using biolistics has met success in fish and marine invertebrates [63] and insects [18]. In practice, this method consists of incubating DNA with gold or platinum. The DNA-coated particles are then launched through the membrane of the embryos.

Hence, only the highly efficient methods of gene transfer may be successful in these conditions even if only random gene integration is needed. The targeted

integration of foreign DNA is a rare event as it is based on a homologous recombination between the host gene and the vector. In these situations, even the most efficient classical techniques of gene transfer are unable to induce a gene targeting at an acceptable rate.

The genetic modifications must then be achieved in cultured cells, which must have the capacity to participate in the development of the embryos and be able to transmit the genetic modification to progeny. Pluripotent cells (ES cells or iPS) may, in the best cases, participate in the generation of chimeric animals in which a significant proportion of the gametes harbor the foreign DNA. Particular multipotent cells, the primordial germ cells (EG cells) in birds can participate in the formation of genetically modified gametes and in the generation of transgenic animals. The animal cloning technique is another possibility. The genetic modification is then performed in somatic cells that are further used to generate transgenic clones by nuclear transfer.

These different methods to generate transgenic animals are schematically represented in Fig. 3.

Mechanisms of Gene Integration

Random Integration The foreign DNA introduced into the cytoplasm by any method recombines according to a random process, leading to the formation of multimers known as concatemers including gene rearrangement and mutations. The different copies of the gene are then organized in head to tail or in tandem. When the foreign DNA is introduced directly into the nucleus (essentially by microinjection), a polymerization also occurs but following a process of homologous recombination generating concatemers organized essentially in tandem with limited mutations and rearrangements. The integration of foreign DNA fragment into a genome may occur by two different mechanisms. The most frequent process is considered as leading to a random integration. Targeted integration is much less frequent.

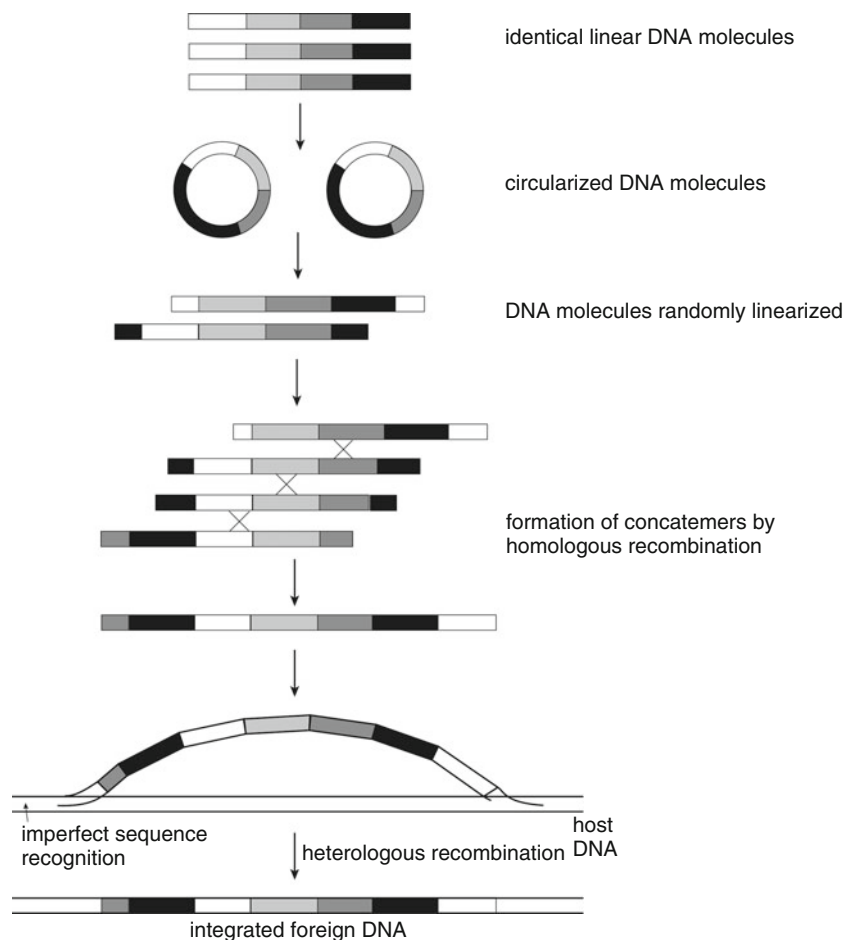
In both cases, the foreign DNA is progressively degraded by the cell. This includes the action of exonucleases, which generate randomly single-stranded sequences in both ends of the foreign DNA. These sequences can recognize complementary regions in the genome, leading to the formation of hybrids. During the cellular DNA replication, these abnormal

structures are corrected by the repair mechanisms of the cell. This leads to the elimination of the foreign DNA or to its integration in the genome [3] (Fig. 4). This process implementing a heterologous recombination is considered as random since the recognition of the host DNA site depends on the action of the exonucleases. The different lines of transgenic animals obtained in this way are thus all different from each other. They contain variable copy number of the transgene integrated each time at a different site. The integration of the foreign gene may damage locally the host DNA. Moreover, the transgene is then frequently submitted to the unpredictable and unknown effects of the transcription regulatory elements

located in its vicinity. The regulatory elements of the transgene may also alter the transcription of the host genes in its vicinity.

The integration of the foreign DNA by these methods is thus not controlled but likely not completely random as it occurs more frequently in regions containing genes and thus having an open chromatin structure, favoring the access of the foreign DNA to the host DNA. The integration site may be known by sequencing the flanking regions of the transgene.

A systematic study revealed that transgenic mice heterozygous for the transgene show rarely abnormalities. On the contrary, the homozygous mice appear



Transgenics: Alternative Gene Transfer Methods. Figure 4

Mechanism of random DNA integration. The ends of the foreign DNA are partially digested generating single-stranded sequences, which randomly recognize complementary host DNA sequences and provoke the integration of the foreign DNA

altered in a proportion as high as 3–10%, suggesting that the uncontrolled integration of the foreign DNA is often mutagenic [103, 104].

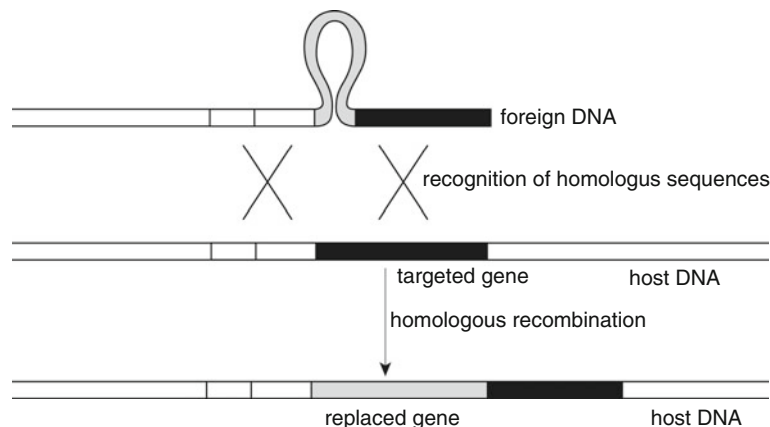
Targeted Integration To avoid the side effects of the random integration, it is possible to target the integration of the foreign gene using homologous recombination. This mechanism is based on the recognition between a genome sequence and the sequence of the exogenous DNA. This recognition leads to the formation of hybrid and finally by the precise replacement of the endogenous gene by the exogenous DNA (Fig. 5). Homologous recombination exists in all living organisms. It is implemented to repair mutated genes using the other allele as a matrix, to redistribute the regions of homologous chromosomes during the formation of gametes, and to generate functional antibody genes from parent genes. Homologous recombination is routinely used to genetically modify bacteria and yeast. Homologous recombination is a rare event corresponding to about 0.1–1% of the heterologous recombination. It is therefore not implemented directly in early embryos but in intermediate cells further used to generate transgenic animals.

Homologous recombination allows in practice the replacement of a given gene by an exogenous DNA fragment. Several applications of this approach are

possible: (1) the replacement of a gene by a non-functional DNA sequence, leading to the inactivation of the targeted gene known as gene knockout (Fig. 5), (2) the targeting of a foreign gene into a given region of the genome or the replacement of an allele by another allele known as knockin.

DNA Microinjection

About 1,000 copies of the isolated foreign gene contained in 1–2 pl may be injected into one of the pronuclei of 1-day mammal embryos. This method implies a superovulation of the females followed by a mating with a male to obtain an optimum number of embryos. The resulting one-cell embryos are collected the next day and microinjected with DNA. The embryos are then transferred to hormonally prepared recipient females using surgery operations. The yield of this method in mice is of 1–3 of transgenics for 100 microinjected and transferred embryos. This rate is relatively low, and this is due to a large extent to the fact that only 50% of the embryos survive after the microinjection. This is clearly due to the mutagenic effect of DNA as the survival of embryos is high when the buffer without DNA is injected. In fact, 1,000 copies of the gene correspond to a very large excess of DNA in the cell, making the multiple random deleterious



Transgenics: Alternative Gene Transfer Methods. Figure 5

Mechanism of targeted integration. An exogenous DNA having a sequence of several kb similar to that of a host gene may replace the host gene using a homologous recombination process. If a non-homologous sequence is surrounded by two homologous sequences, the two independent homologous recombinations direct the integration of the non-homologous sequence, leading to the knockout of the targeted gene

integrations and genome rearrangements depicted above possible. Despite its drawback and the fact that it is laborious, this technique is still the most frequently used in mice and rabbits. The efficiency of DNA microinjection is lower in all the other mammalian species and very low in ruminants. This is not due to the difficulty to inject DNA but to the fact that the mechanism of integration is for unknown reasons much less efficient in some species.

The DNA microinjection in pronuclei gives birth to at least 30% of mice mosaic for the transgene. The transgene is then not present in all the cells of the transgenic founder. This is due to the fact that the integration of the foreign DNA occurs sometimes not in the first cell but later at the two- or four-cell stage. The transmission of the transgene from these founders appears not to respect the Mendel law. The transmission rate is due to the fact that the transgene is not present in all the gametes. At the next generation, the proportion of transgenics is Mendelian [24]. About 1% of transgenic founders do not transmit their transgene as the mosaicism is very high and the transgene is rare or non-existent in gametes.

In non-mammalian species, the pronuclei cannot be visualized as the embryo is embedded in an abundant and opaque vitellus. High amounts of DNA (millions of copies in a few nanoliters) must then be injected into the cytoplasm of the one-cell embryos. This relatively simple technique is efficient in several fish species [42, 56] but highly inefficient in chicken, in *Xenopus*, in some fish, and in insects. For unknown reasons, the integration of the foreign DNA thus does not occur in some species. In lower vertebrates and invertebrates, the embryo DNA and the foreign DNA are replicated very rapidly. The foreign DNA is then highly amplified leading to multiple independent integrations occurring during the first days of embryo development in the same animals. The resulting transgenic animals are often very mosaic making it difficult their use. Several reproduction cycles allow a segregation of the different transgenes until the animals contain a single integration site.

In a *C. elegans*, direct DNA injection into gonad syncytium leads to the generation of transgenics with a good yield [98].

Gene constructs bordered by I-SceI sites have been used to generate transgenic fish. The plasmids were

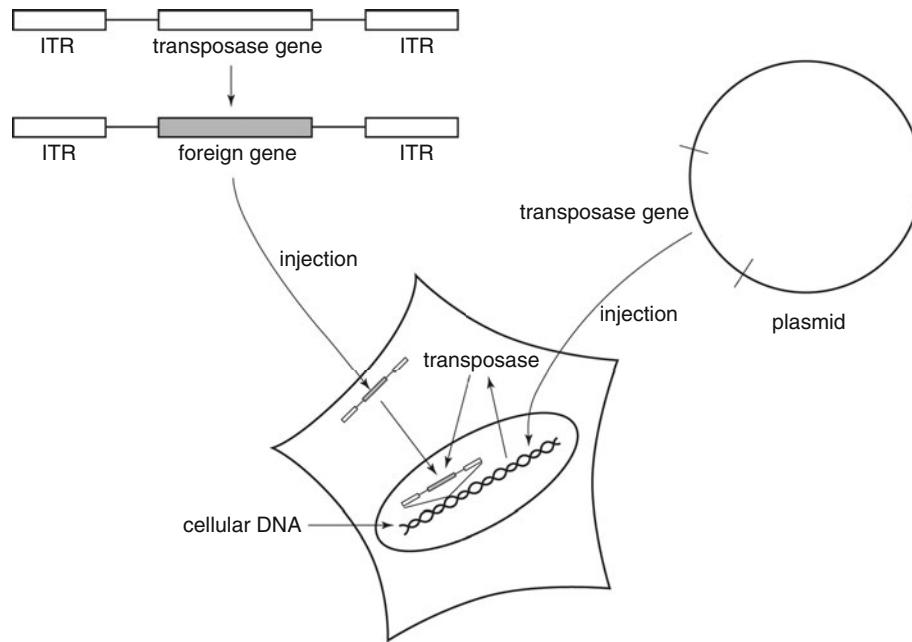
previously cut by the I-SceI, and the fraction was injected without any previous DNA purification. Unexpectedly, this protocol markedly increased the yield of transgenesis. Interestingly, the same observation was made in *Xenopus*. The mechanism improving transgenesis is not known [97]. The I-SceI might protect DNA against degradation, favor DNA transfer to the nucleus, or directly stimulate the integration mechanism.

Microinjection is therefore a good technique but insufficient in some species. Additional methods have been found and are still under study.

Detection of the transgene and examination of its integrity can be achieved using Southern blotting or PCR using multiple primers corresponding to various regions of the transgene. Zygosity of the transgene may be evaluated by quantitative PCR [69] or in situ hybridization.

Use of Transposons

Transposons are short genomic DNA regions, which are replicated and randomly integrated into the same genome. The number of a given transposon is thus increasing until the cell blocks this process to protect itself from the degradation of its genome. Foreign genes can be introduced into transposons in vitro. The recombinant transposons may then be microinjected into 1-day embryos as it is for DNA. A transposon contains a gene coding for an integrase specific of this type of transposon and required for the integration of the transposon. The integrase gene is bordered by two inverted terminal repeat regions (ITR) required for the transposon integration. The addition of the foreign DNA implies the elimination of the integrase gene. This makes space for the foreign DNA, and it also prevents the recombinant transposon to disseminate in the genome. To allow the integration of the recombinant transposon, the integrase protein or a gene construction directing the synthesis of the integrase must be co-injected with the recombinant transposon (Fig. 6). The foreign gene thus becomes integrated into the embryos with a yield of about 1%. Essentially, all the transgenic insects are generated by using transposons as vectors [43, 95]. Transposons also proved efficient to generate transgenic fish, chicken, and mammals [22, 23]. The transposons used for transgenesis



Transgenics: Alternative Gene Transfer Methods. Figure 6

Gene transfer using a transposon. The integrase gene of a transposon is replaced by the gene of interest between the two ITR (inverted terminal repeats) required with the integrase for an efficient integration. Integrase or a gene coding for this protein must be injected with the transposon to allow its integration

are chosen or engineered to be unable to disseminate in the host genome under the action of the integrase of endogenous transposons. Transposons are therefore efficient and safe tools, but they can harbor no more than 2–3 kb of foreign DNA.

Transposons are being used with ICSI giving a high transgenesis rate [64].

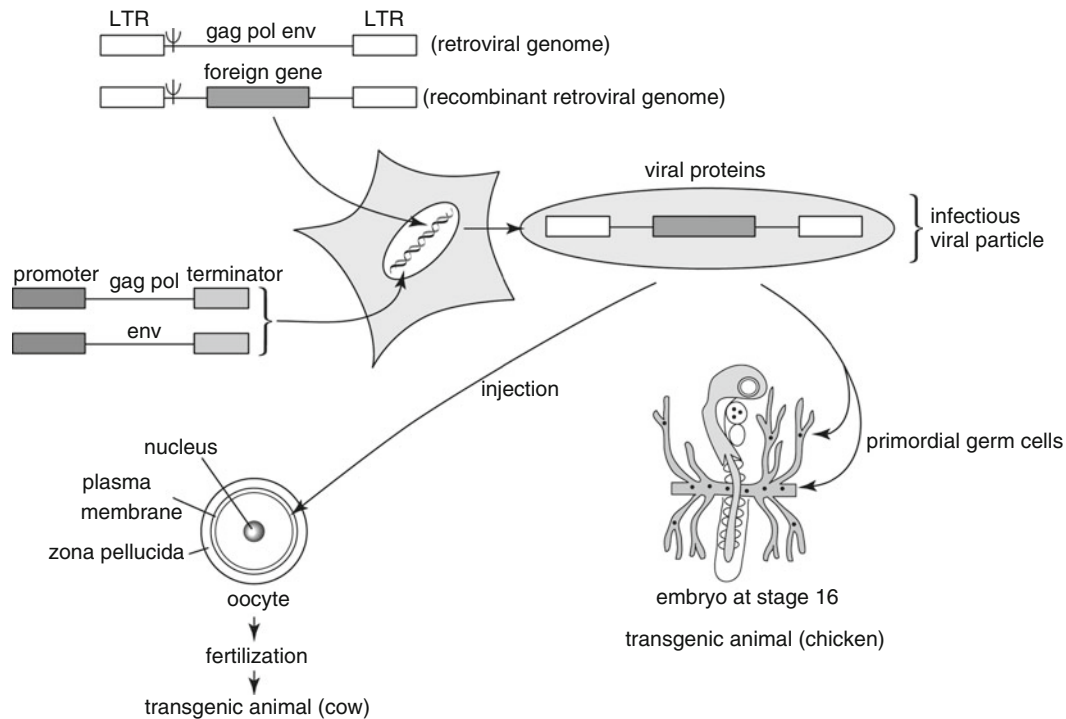
Efficient transposons started being used to disseminate in genomes after the addition of the corresponding integrase into cells. This leads to multiple random integrations of the transposon providing researchers with insertional mutants. A correlation may then be established between a biological alteration of the animals and the nature of the gene inactivated by the presence of the transposon within the gene [22].

Use of Lentiviral Vectors

Retroviruses have not the capacity to autoreplicate, and they must be integrated stably into the genome of the cells they infected to replicate. This explains why up to 1% of animal genomes contain degenerated retroviral

genes. This property of retroviruses is implemented to integrate foreign genes in cells, in animals, and in patient's somatic cells. For this purpose, the viral genes are removed from the genome of lentiviruses (a category of retroviruses) and replaced by the genes of interest. Viral particles are then prepared and used to transfer the foreign genes into oocytes or one-cell embryos (Fig. 7). Retroviral vectors have been studied and used during the last two decades for gene therapy. Important improvements of these vectors have markedly enhanced their efficiency. Vectors using ALV (avian leucosis virus) studied two decades ago proved poorly efficient to generate transgenic chicken [4, 83].

The lentiviruses have the advantage over the common retroviruses to be able to infect and transfer their genes in quiescent as well as in multiplying cells. The majority of the retrovirus genomes need a cell multiplication including a transient degradation of the nuclear membrane to reach chromatin and integrate into the cell genome. On the contrary, lentiviral viruses contain proteins carrying their genome to the nucleus in quiescent as well as in multiplying cells.



Transgenics: Alternative Gene Transfer Methods. Figure 7

Gene transfer using lentiviral vectors. The pathogenic genes are removed from the HIV genome and replaced by the gene of interest. The gene construction is transfected into trans-complementing cells synthesizing the essential HIV proteins required for the formation of infectious recombinant particles. The particles secreted from the cells in the culture medium are injected between the zona pellucida and the membrane of the cell embryos or of oocytes or within the fetal gonad in chicken. The infection is followed by an efficient integration of the gene of interest

EIAV (equine infectious anemia virus) proved that it is able to generate transgenic chicken [60]. Presently, the most frequently used lentiviral vector derives from the HIV (human immunodeficiency virus) genome. The natural envelope protein of the HIV, which recognizes specific receptors on the plasma membrane to induce infection has been replaced by the envelope of vesicular stomatitis virus, which is not a retrovirus. This envelope recognizes the phospholipids of plasma membrane, making the infection of essentially all cell types possible. The particles containing this envelope are stable, and they can be concentrated and administered at a high concentration into embryos or oocytes. For unknown reasons, lentiviral vectors have to be injected in cow oocytes rather than in one-cell embryos to transfer their genes.

Safe experimental conditions have been defined to use the lentiviral vectors. This method proved very

highly efficient in several species including mammals [74, 78] and birds [11, 52, 86]. Up to 90% of transgenic animals may be obtained. This results from the fact that with high viral particle concentration, multiple independent integrations take place in the embryo genome. These integrations may occur later than the first cell stage giving birth to mosaic animals. The selection of lines harboring a single integration is preferable, but it may take a very long time in farm animals. Reducing the concentration of viral particles reduces the yield of transgenics but also the multiple integrations.

Transgenes transferred by lentiviral vectors are reputed not to become silent on the contrary to transgenes transferred by plasmidic constructions. This may be due in part to the fact that one or several integrated copies are active whereas others have been silenced. Lentiviral proved to be efficient tools to express transgenes coding for siRNAs [99].

Interestingly, siRNA could induce specific gene knock-down in rat [19, 35]. This observation is important as the classical tools to induce gene knockout (ES cells and cloning) are not available in this species. Lentiviral harboring foreign genes driven by cell-specific promoters may direct the expression of the transgenes in the targeted cells specifically [86]. The use of lentiviral vectors is limited by the fact that they cannot harbor more than 8 kb of foreign DNA. Special constructions must also sometimes be prepared to reduce the influence of the viral enhancers on the transgene promoter.

In mammals, the lentiviral vectors are injected between the zona pellucida and the cell membrane of one-cell embryos or oocytes. The efficiency of lentiviral vectors may be enhanced by injecting envelope-free constructions into cell cytoplasm [108].

The preparation of lentiviral particles carrying foreign genes is routinely achieved by academic and private structures at a moderate price. This possibility is particularly attractive for laboratories, which do not need a frequent use of this tool.

Use of ICSI

Foreign DNA must not necessarily be introduced into one-cell embryos. In principle, it can be as well injected in gametes before fertilization. DNA microinjection into oocytes proved inefficient, and this approach was soon abandoned. Using sperm is another possibility based on the assumption that foreign DNA cannot integrate into the genome as in this cell DNA is embedded in protamines and not able to replicate. Sperm was rather postulated to bind DNA in its surface and carry it into the oocyte during fertilization.

More than a decade ago, it was shown that sperm incubated in the presence of DNA before being used for fertilization was able to transfer the foreign gene into the oocyte and generate transgenic mice. This approach proved capable of generating transgenic mice, sheep, chicken, fish, and pigs [45, 109]. However, this method appeared difficult to use as DNA was frequently degraded [91]. The seminal plasma contains high DNase activity, which degrades the gene and only rearranged fragments of the foreign genes were generally found in the transgenic animals. An appropriate sperm washing is required to prevent DNA degradation. For no clear reasons, not all the ejaculates can lead

to a success of the method. Transgenic mice and rabbits were obtained by incubating sperm with DNA in the presence of DMSO (dimethylsulfoxide) and by using conventional *in vitro* fertilization [87].

The method has been greatly improved by using ICSI (intracytoplasmic Sperm Injection). This technique, which consists of injecting sperm into the cytoplasm of oocytes is currently used for *in vitro* fertilization in humans. To transfer genes, sperms from which plasma membrane has been damaged by freezing and thawing or with a mild treatment by non-polar detergents were incubated in the presence of the gene of interest and further used for fertilization by ICSI. The ICSI approach started being implemented in *Xenopus* a decade ago as all the other methods failed [59]. This method proved efficient in mice [65, 88] and pigs [109]. Transposon use and ICSI may be combined to increase the yield of transgenesis [64, 88].

ICSI is therefore an excellent method to generate transgenic animals on condition that ICSI is possible in the considered species. One advantage of ICSI is that both long and short DNA fragments of DNA may be used successfully. DNA rearrangements may occur during ICSI especially when long DNA fragments are used. Another advantage is that foreign DNA is integrated mostly at the first cell stage of embryos. This reduces the number of animals being mosaic for the transgene.

Use of Episomal Vectors

The methods described above to transfer foreign genes rely on the integration of the DNA into the host genome. Another theoretical possibility is the use of episomal vectors capable of autoreplicating in host cells independently of the genomic DNA and transferred to daughter cells. Fragments of chromosomes are being used for particular projects, requiring the transfer of very long DNA fragments [44]. These chromosomal vectors are not of an easy use, and they carry a number of genes in addition to the gene of interest. These extra genes may interfere with the transgene or with the whole organism of the host.

Another possibility consists of using vectors, which derive from viruses having the capacity to replicate in animal cells and to be transferred to daughter cells. Herpes viruses are naturally stably maintained as autonomous circular minichromosome at a low copy

number in animal cells. Foreign genes can be introduced into Herpes viral vectors and be maintained during cell division. This kind of vectors is generally species specific. This greatly reduces their potential use as well-known Herpes viruses are not available for all animal species.

Episomal vectors not based on the use of viral elements are available. Such a vector proved highly efficient to transfer foreign genes into pig embryo using ICSI [58]. This vector was maintained without any selection pressure in the cells of the early developing embryos but seemingly not later. These vectors are therefore excellent tools to study transgene effect during early embryo development. Hence, until now, only the integration of foreign DNA into the host genome makes it possible the generation of stable lines of transgenic animals.

A theoretical safety problem raised by the use of episomal vectors is that it could be transmitted to other cells mechanically and thus independently of any infectious process. However, this event is expected to occur very rarely. More likely, non-integrated DNA is more available to recombine with homologous sequences in the host genome, leading to mutations, chromosome rearrangement, or integration.

Use of Intermediate Cells

Use of Pluripotent Cells In some situations, the efficiency of the genetic modification is too low to be achieved by the methods described above. This is particularly the case for gene targeting. One possibility to circumvent this problem is to genetically modify pluripotent cells that are further used to participate in the development of living organisms. Pluripotent cells have the capacity to participate in the development of all the organs including gametes.

Pluripotent cells exist in the early embryos (morula and blastocysts), and they are known as ES cells (embryonic stem cells). The pluripotent cells can be cultured, genetically modified, selected and transferred into morula or blastocysts. These cells participate in the development of the embryo to give birth to chimeric transgenic animals (Fig. 3). This means that the organs of the animals, including sexual cells, derive from the genetically modified cells or from the recipient embryo. A proportion of the offsprings from these chimeric

animals harbor the genetic modification if they derive from the transplanted cells.

The first ES cells implemented to genetically modify animals (mice) were used at the end of the 1980s. For unknown reasons, commonly used ES cell lines have been established only in two mouse lines and mainly in the 129/Sv strain. In other lines and species, the ES cells lose their pluripotency. They can give birth to animals, which are still chimeric but have no more the capacity to transmit the genetic modification to their offspring. In practice, this complicates sometimes the use of the knockout mice. Indeed, in a number of cases, the gene knockout should be obtained in a genetic background very different from that of the 129/Sv strain. Successive crossings allow the mutation to be transferred in the strain in which the biological function in question may be optimally studied.

After sustained efforts for two decades, authentic rat ES cell lines have been established [7, 2, 49–51]. This breakthrough relies on the use of small molecules added into the culture medium. These molecules control essential genes required for the maintenance of pluripotency. They replace proteins like LIF, which are commonly used for mouse ES cells but are not successful in other species.

Recent experiments have shown that the transfer of three genes, normally expressed in pluripotent cells, into somatic cells can dedifferentiate these organ cells into pluripotent cells known as iPS (induced pluripotent cells) and almost similar to ES cells [68, 72, 93, 105]. Interestingly, small molecules can mimic and replace Klf4 gene, which is one of the genes required to transform somatic cells into pluripotent cells [55]. These experiments open avenues for cell therapy and gene therapy [76]. The approach known as therapeutic cloning and based on the capacity of cloning (SCNT) to dedifferentiate somatic cells into totipotent cells further differentiated in vitro into pluripotent cells has become virtually no more strictly necessary. iPS cells can potentially be obtained in different species by this method. Similarly, iPS cells might be implemented for transgenesis in species in which ES cells are not available (Fig. 3). Cow and pig iPS cells have been obtained and are currently under study.

Use of Primordial Germ Cells and Testis Stem Cells Experiments carried out a few years ago showed that

chicken primordial germ cells (PGC) can be isolated and cultured in conditions maintaining their multipotency giving stable cell lines known as EG cells (embryonic gonad cells). Foreign genes can be transferred into EG cells, which can be implanted into recipient embryos and participate in gonad development. In practice, the EG cells, which contain the gene of interest and a selection gene are cloned, amplified, and injected into an early embryo in which the majority of the cells have been destroyed by irradiation. This gives the best chance to the EG cells to colonize the embryo and to give birth to transgenic showing a high degree of chimerism and thus transmitting their transgene to progeny with a high yield. This approach has greatly simplified the generation of transgenic chicken [33, 101].

Testicular cells, which are sperm precursors can be isolated, cultured, genetically modified, more or less differentiated *in vitro*, and transplanted into recipient testis to give functional sperms that are able to generate transgenic animals by fertilization. Alternatively, sperm cell precursors may be genetically modified *in situ* using viral vectors [33, 41, 46, 48, 94]. These methods are still under study, and they are not currently used to generate transgenic animals.

Use of Cloning

The birth of Dolly, the sheep, demonstrated that the genome of somatic cells can be reprogrammed after being introduced into an enucleated oocyte. This generates a pseudo-embryo capable, with a relatively low yield, of giving birth to clones of the cell donor. This technique was initially designed to improve transgenesis efficiency in farm animals. This approach is likely to be used to accelerate genetic selection, but its only real application is presently transgenesis [81, 85]. The principle of this method is described in Fig. 3, and this topic is the matter of another chapter. Genes are transferred into somatic cells, which are then used to generate transgenic clones. This method has become the most frequently used for big farm animals as it simplifies the task of experimenters and enhances the rate of transgenic animals. Recently published important data have shown that the cloning technique does not provoke mutations in the clones [67].

Gene Construction

Random Integration A problem, which has not been completely solved is the reliability of transgene expression [36–38]. In the early 1980s, the first experiments to generate transgenic mice revealed that transgenes were often not working as expected. In a number of cases, the expression of the transgenes was very weak and not strictly specific to the promoter associated with the foreign gene. In a few cases, it was demonstrated that the ectopic expression of the transgenes was due to the presence of genomic enhancers in the vicinity of the integrated foreign DNA. The frequent transgene silencing was thought to be induced by the integration of the foreign genes near genomic silencers. These putative silencers were rarely identified suggesting that the ectopic transgene expression and their silencing could be not symmetrical phenomena. It was also proved that the level of transgene expression was generally not a function of the integrated copy number. In a number of cases, the expression level appeared even lower when the number of integrated copies was higher. A striking demonstration was given by the experiment in which the human β -globin gene was bordered by two LoxP sequences and integrated into mouse genome as several copies in tandem. The transgene remained silent in these mice but was reactivated in their offspring in which the copy number was reduced to one by the action of the Cre recombinase [26].

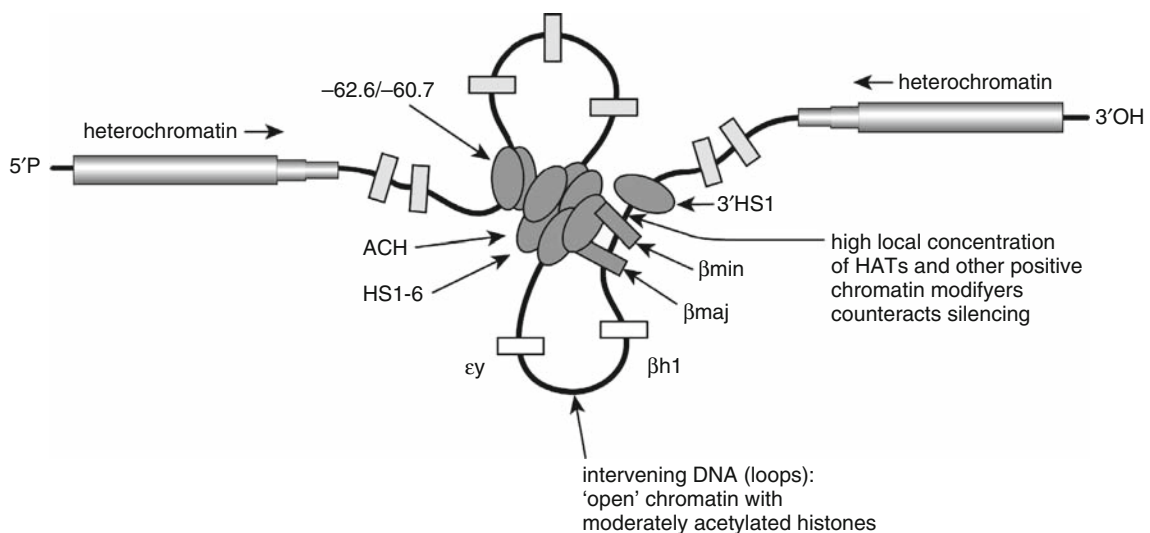
After about one decade, it appeared that this was due to chromatin position effects suggesting that the transgenes were recognized as foreign sequences by some unknown cellular mechanisms. One of the most surprising data was that a genomic DNA sequence containing the whole human β -globin gene including its promoter region allowing the gene to be expressed as expected in cultured red blood cells remained silent in transgenic mice. This discrepancy suggested that the transgene silencing occurred more *in vivo* than in cultured cells and that this phenomenon could take place during the early phase of embryo development. A hypothesis was also that the genomic DNA sequence contained the whole β -globin gene and some but not all the transcription regulators. A confrontation of the very low expression level in patients suffering from β -thalassemia and the structure of their DNA in the genomic β -globin gene region revealed that, in some

cases, the gene and its promoter were normal but that some remote genome regions were missing. This suggested that these regions could be the putative regulators missing in transgenic mice. An association of these regions with the β -globin gene allowed the later to be highly expressed in transgenic mice. The extensive study of the β -globin gene locus in several species revealed that remote regulatory elements are present on both side of the locus. These elements bind transcription factors specifically present in differentiated red blood cells, and they form a transcription complex known as a hub in the vicinity of the promoter through a looping process (Fig. 8) ([20, 21]). This type of mechanism seems to be common to many if not all genes in vertebrates and similarly also in invertebrates.

These observations may explain at least in part why traditionally constructed transgenes are so often poorly active and they suggest using long genomic DNA fragments contained in BAC (bacterial artificial chromosome) vectors to promote transgene expression [53]. The implementation of the long DNA fragments may be laborious as they are sensitive to mechanical degradation [100, 102]. The long DNA fragments can be transferred using microinjection, ICSI, gene targeting, or via intermediate cells. The failure of transgene activity is likely due to multiple reasons, which still

complicate the construction of vectors allowing an efficient and reliable expression of transgenes. The optimized conditions to use long DNA fragments as vectors are described in a recent paper [102].

Nucleotidic Composition of the Vectors Integrated retroviral sequences and transposons are inactivated by a cytosine methylation of the CpG motifs and the local formation of condensed chromatin (heterochromatin) in which histones are deacetylated and methylated in some sites. Transgenes seem to be inactivated by similar mechanisms. Many of the vertebrate genes contain CpG islets in their regulatory regions, which contribute to their expression. Some of the CpG motifs belong to the binding site of the transcription factor Sp1, which is present not only in the promoter region of the gene but also sometimes in the first introns. An exceedingly large number CpG motif in vectors induces transgene silencing. The replacement of some of the GC-rich regions by AT-rich regions improves transgene expression. MARs (matrix-attached region) are frequently found in the vicinity of genes, and they bind DNA to the nuclear matrix locally. MARs are generally AT rich, and they have been added into vectors to tentatively improve transgene expression. This approach met variable success. The *Escherichia coli* β -galactosidase gene is rich in



Transgenics: Alternative Gene Transfer Methods. Figure 8

Mechanism of action of remote transcription enhancers. The formation of loops allows the various regulatory factors to be in close vicinity and generate an active transcription complex

CpG, and it is known to be a potent transgene silencer. This silencing potency proved to be markedly reduced as the number of CpG was diminished [17, 34]. The coding sequences of a transgene may thus be obtained by chemical synthesis to replace a part of the CpG-rich codons by others without modifying the sequence of the corresponding protein.

Use of Insulators In order to improve the expression of transgenes, it is possible to use large genomic DNA fragments (50–250 kb) expected to contain all the regulatory elements of the gene of interest [53]. It implies only the isolation and characterization of BACs (bacterial artificial chromosome) from a bank. Linearized or circular BACs may be used as such if the expression of the gene(s) they contain is wanted [102]. One drawback of using BACs is that they often contain several genes. The BACs thus transfer all these genes potentially generating unknown and unwanted interactions with the animals. If needed, these genes may be inactivated in BACs by performing short deletions, for example, of the cap region, using homologous recombination in bacteria.

An attractive approach consists of using BACs as vectors harboring the foreign genes. The foreign DNA sequence must be introduced into the BAC using homologous recombination in bacteria. It is important to note that the transgenes driven by BACs rarely work in an ideal fashion, if only this concept has a real meaning. Long genomic DNA fragments are expected to suppress the position effects, which is not often fully the case. Indeed, it is clear that the variegated expression, which characterizes the conventional transgenes is less or even much less frequent in animals harboring BAC vectors. A higher proportion of animals expressing the transgenes is generally found with BAC than with plasmid vectors. Yet, the different lines harboring BACs vectors usually do not express the transgenes at an identical level for a given number of integrated copies. Similarly, it has been rarely reported that the expression of the transgenes was strictly a function of the copy number of the integrated BACs. This means that the BACs provide transgenes with essential elements for their expression but they remain often not fully able to suppress the position effects. This disappointing observation may not be surprising from a theoretical point of view.

Indeed, a locus has been constructed during evolution to express in an appropriate manner the genes it contains. This implies that these genes are protected against deleterious position effects in their natural chromatin environment. They have no theoretical reasons to be independent of the position effect in all their integration sites. Some BACs may contain all the elements providing transgenes with a complete independence of the integration site. If not, a BAC vector may still contain enough regulatory elements improving significantly transgene expression to justify its use.

A more sophisticated approach could consist of using as vectors containing not all the DNA sequence of BACs but only the major elements involved in the control of gene and transgene expression. This implies that these regulatory elements have been identified, characterized, and introduced into mini-BACs or even plasmids. An example of this is a major regulatory region of the genes present in mammalian β -globin locus, which is located upstream within the locus of olfactory receptor genes [21]. The study of the remote genomic regulatory elements is still in infancy. Apart from the technical difficulty to study them is the fact that each gene or gene cluster seems to have used available genomic sequences to design specific mechanisms allowing a satisfactory expression. Some of these regulatory elements have an unexpected structure. Examples are SINE B2 and Alu sequences or some active tRNA genes, which are essential regulators for neighbor genes [54].

The notion of boundary elements and insulators is essential to understand how unrelated genes can be expressed in a specific manner without being under the dependency of the neighbor gene regulators. This situation is clearly often not encountered for transgenes suggesting that the constructs commonly used do not contain the natural insulators. Insulator activity has been found in the LCR (locus control region) of the chicken β -globin locus. This activity was identified in a 300 bp fragment, which proved its ability to block in a specific sense the action of an enhancer when added between this enhancer and a promoter directing the expression of a reporter gene. The enhancer blocker was mediated by the binding of the regulatory protein CTCF to a specific DNA sequence. The CTCF element has now been found in the boundary region of several other genes. This type of elements, known as enhancer-blocking

insulators, cannot be assimilated to silencers as the former act only when they are located between an enhancer and a promoter. Moreover, the enhancer-blocking insulators often show unidirectional action.

This 300-bp region of the chicken β -globin locus was found later to contain another sequence known as a chromatin opener. Chromatin openers are regulators capable of maintaining a local euchromatin configuration favoring the expression of the neighbor gene by preventing the local formation of condensed chromatin (heterochromatin) [27]. The elements having this function are known as barrier insulators [27]. The barrier insulators cannot be assimilated to enhancers as their effect does not occur during transient foreign gene expression in transfected cells. The region locus known as 5'HS4 and containing the 300-bp sequence can improve the expression of a number of unrelated transgenes in mammals [29, 92]. However, the potency of the 5'HS4 element remains generally insufficient to express transgenes in a fully satisfactory manner.

In the mean time, the presence of AT-rich MARs within the genomic region required for the expression of a gene suggested that these sequences were essential remote regulators. This hypothesis was not confirmed [90].

Optimization of the Transcribed Region The 5'UTR (untranslated region) must have less GC sequences stabilizing double-stranded hairpin structures, which do not favor ribosome migration to the initiation codon. The AUG initiation codon must preferably be in the Kozak consensus sequence GCCA/GCCAUGG to optimize translation initiation. The natural 5'UTR of the gene of interest may contain sequences regulating translation. It may be then useful not to keep this region and replace it by a short (not less than 80 nucleotides) AT-rich 5'UTR region from gene known to be efficiently translated in many cell types or in the targeted cells of the animals. Some mRNAs encode proteins, which are not naturally secreted. Peptide signals may be added to their cDNA.

A transgene must contain at least one intron, which is required to favor the transfer of the mRNA to the cytoplasm. The first intron of many genes contains sites, which bind transcription factors and they may favor transgene expression. The intron excision is dependent upon several signals comprising consensus

sequences in both splicing sites (CAG GUA/GAGUA/UGGG in 5' and CAG G...GAA/G...GAA/G...in 3'), namely, a CU-rich region immediately upstream of the 3'splicing site and a BPS site (branched point sequence) U/CNCUGAC at about 30 nucleotides upstream of the 3'splicing site and upon splicing enhancers [61]. The second intron of the rabbit β -globin gene is considered as efficient to express transgenes in mammals. The intron(s) must preferably be put before the coding region. If an intron is added after the translated region, the 5'splicing site must be located not more than 50 nucleotides from the termination codon to avoid the activation of the NMD (nonsense-mediated decay), which degrades the mRNA [10].

The cDNA and other regions of the vectors may preferably be chemically synthesized. This allows reducing the number of CpG motifs, to choose the best codons, to eliminate cryptic 3' or 5'splicing sites and sequences known to prevent transcription or translation.

The 3'UTR region of a number of mRNA contains signals for mRNA translation and stability. A number of mRNAs have an AU-rich region with the AUUUA motif in their 3'UTR. These mRNAs have a short half-life controlled by the cell cycle (Beelman and Parker 1995). The fortuitous presence of such sequences must be searched and eliminated to prevent a poor transgene expression. Some mRNAs contain translation regulators acting by the binding of proteins favoring the recycling of ribosomes by binding to the 5'UTR. CU-rich regions in the 3'UTR enhance the stability of the mRNAs, and they may be added in the vectors downstream of the cDNAs. Stabilizing sequence can be taken in the 3'UTR of the human or bovine genes and of the α -globin gene, which also contain efficient transcription terminators [13]. Some proteins are anchored to the plasma membrane by a GPI structure (glycophosphatidylinositol). A protein normally not anchored in this way acquires this property by adding in the 3'end of the cDNA the peptide, allowing the addition of GPI. Micro-RNAs (miRNA), the role of which was recently discovered, inhibit specifically the translation of an mRNA after forming a hybrid with its 3'UTR. The presence of target sequence for a miRNA may unduly inhibit the expression of a transgene. This target sequence should then be deleted.

Targeted Integration The techniques described above lead to uncontrolled but not strictly random gene integration. Foreign DNA is preferentially integrated into gene-rich genome regions, and its location can be precisely identified. A foreign DNA fragment can recombine very precisely with a genomic DNA region containing a similar sequence. This natural mechanism known as homologous recombination makes the precise replacement of a gene by another possible. An active gene may thus be replaced by an inactive version, leading precisely to an inactivation of the targeted gene (gene knockout).

The targeted gene may as well be replaced by an active gene (gene knockin). This technique allows therefore a better controlled transgenesis reducing possible damage of the genomic DNA at the integration site and frequent side effects of the genes located in the vicinity of the transgene on the expression of the transgene. Two mouse genomic loci are currently being used as foreign gene knockin. One is the HPRT locus [6] and the other is Rosa 26 locus [110]. The genes of these loci are known to be expressed constitutively, and they were supposed to be bordered by regulators able to drive the expression of transgenes in a reliable manner. In practice, a number of transgenes appeared to be expressed as expected when integrated in these loci. Interestingly, also the expression of the transgenes added in these loci remained specific to the promoter linked to the foreign genes. The two loci thus appear to maintain an open chromatin configuration, favoring the expression of the transgenes irrespective of their composition.

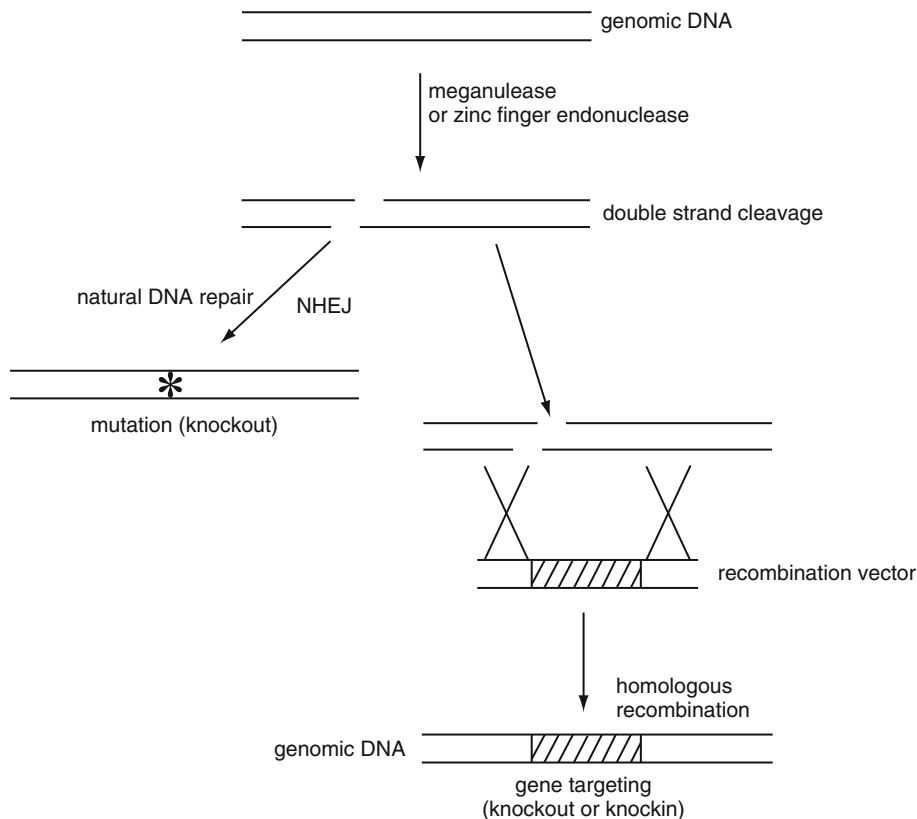
Yet, this approach remains limited by the fact that the homologous recombination required for gene targeting is a rare event. The targeted integrations by homologous recombination of a foreign DNA represent 0.1–1% of the total integrations. The cells in which targeted integration occurred must be selected and used to generate a transgenic animal. The formation of chimeric embryos using pluripotent cells, multipotent cells, or the cloning technique is presently required to obtain a targeted integration.

Meganucleases were discovered in yeast about two decades ago, and it was shown that they participate in intron generation. These enzymes recognize sites as long as 18 nucleotides leaving negligible chance to cleave mammalian genomic DNA. About 80 natural meganucleases have been identified so far. It was

observed later than one of these meganucleases I-SceI induces chromosome recombination in mammalian cells and this was attributed to the fact that meganucleases cleave both DNA strands [14]. Such breaks stimulate DNA repair mechanism. It was also shown that I-SceI amplifies the rate of homologous recombination [15] and consequently gene replacement in mammalian cells [16]. The DNA sequences recognized by natural meganucleases must then be added to the genome of animals either at targeted sites by homologous recombination or at random sites. To circumvent this problem, hundreds of engineered meganucleases each recognizing a specific genomic DNA sites have been obtained making a local cleavage of DNA and gene targeting possible [25].

Multiple engineered zinc finger nucleases (ZFN) have also been obtained. Each of these enzymes has the capacity to cut one DNA strand at a unique genomic DNA site. To mimic natural meganucleases and cut locally both DNA strands, two ZFN each recognizing a DNA strand at neighbor sites are needed [75]. Engineered meganucleases and engineered zinc finger nucleases, thus, make it possible gene targeting at multiple sites of the genome. This method, which is being developed to improve the efficiency and the precision of gene therapy for humans can be applied to target the integration of foreign DNA into experimental animals. Interestingly, when the recombination vector is not added with the meganuclease or the ZFN, the genomic DNA repair takes place spontaneously but often with alteration of the sequence. This process known as NHEJ (non-homologous end joining) corresponds to a knockout (Fig. 9) [84, 106]. This mechanism is efficient, and it allowed a knockout in one cell fish and rat embryos after the injection of engineered meganucleases [28, 107]. This method can be considered as a transgenesis without transgene. This inclines to think that gene targeting might be achieved directly in mammal embryos by injecting an engineered meganuclease or ZFN with or without a homologous recombination vector.

In the same line, the bacterial enzyme phiC31, which is an integrase, recognizes several sites in various animal genomes and allows the efficient integration of foreign genes at the targeted sites. Several other recombination systems rely on the use of integrases such as Cre and Flp, which recognize specific sites of about



Transgenics: Alternative Gene Transfer Methods. Figure 9

Engineered meganucleases or ZFN (zinc finger nucleases) induce a targeted local cleavage of both genomic DNA strands. In the absence of recombination vector, the DNA is repaired leading to frequent local mutations and to a knockout. In the presence of a recombination vector, the foreign gene is integrated into the targeted site with a high efficiency

30 nucleotides (LoxP and FRT, respectively), which must be added to the animal genome. In practice, deletion of the integrated foreign DNA has more chance to occur than integration. Mutated LoxP and FRT sequences capable of promoting integration but not deletion must be used. This approach is known as RMCE (Recombinase-mediated Cassette Exchange) [1]. These systems are more often used to delete a DNA region previously bordered by the LoxP or the FRT sequences.

The Co-expression of Several Cistrons It is sometimes necessary to express two or even three genes in the same transgenic animals. The co-injection of several independent vectors makes the generation of up to 80% of the animals harboring the two or three genes, which are co-integrated at the same site possible.

An alternative consists of using IRES (internal ribosome entry site). Such sequences exist in the 5'UTR of many mRNAs, the translation of which is controlled by these sequences, which bind specific cellular inducible proteins. Such sequences may be added between two cistrons and allow their simultaneous translation from a single vector. The addition of the IRES 80 nucleotides after the termination codon of the first cistron may contribute to favor the expression of the second cistron [39]. It should be kept in mind that IRES represents a family of sequences acting by different mechanisms, which are only partly known.

Gene Inactivation (Knockdown) with Interfering RNAs Long double-stranded RNAs are randomly cut into 19–21 nucleotide fragments known as siRNA (small interfering RNA) or RNAi. One of the two

strands of the siRNA is kept and targeted to an mRNA having a complementary sequence. This induces the degradation of the mRNA or the reversible inhibition of its translation. In practise, a synthetic gene containing the targeted 19–21 nucleotide sequence followed a short random sequence and by the targeted sequence in the opposite orientation is linked to a promoter acting with RNA polymerase III (usually U6 or H1 gene promoters). The RNAs synthesized by such vectors form a 19–21 nucleotide double-stranded RNA known as shRNAs (short hairpin RNA) and are processed in cells to generate active siRNAs. An appropriate expression of siRNA genes in transgenic animals can be obtained when they are introduced into lentiviral vectors [99].

The recent discovery of the role of micro-RNAs has increased the possibility to use interfering RNAs. Micro-RNAs are encoded by short genes expressed under the control of RNA polymerase II promoters. Their primary products are transformed into siRNAs. Alternatively, the sequence coding for a micro-RNA and even several micro-RNAs in tandem may be inserted into introns without any promoter. The promoter of the gene or transgene thus drives transcription of the micro-RNA genes and the intron degradation releases micro-RNA precursors, which are processed in the nucleus and the cytoplasm, generating active siRNAs (Ripoll et al. unpublished data). The mature miRNAs, which are fully complementary to the targeted mRNA induce degradation of this mRNA. The miRNAs, which are only partially complementary to the targeted mRNA and which recognize a sequence located in the 3'UTR (3'untranslated region) of the mRNA inhibit translation of this mRNA without inducing its degradation. The possibility known as knockdown to generate transgenic animals expressing siRNAs preventing specifically the expression of a gene by degrading the corresponding mRNA or inhibiting its translation has opened avenues for the control of gene expression in vivo. The application of the siRNA approach raises specific problems in animals. Long double-stranded RNAs induce interferons and some unspecific immune reactions [89]. On the other hand, siRNAs are not autoamplified in higher animals and this reduces their potency. Vectors to express miRNA gene are available, but simple shRNA genes are not easily expressed in transgenic animals

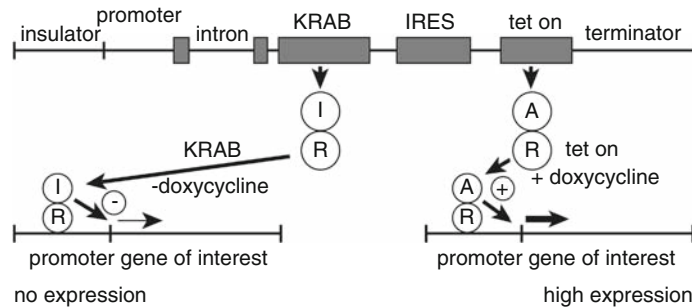
using conventional vectors. Moreover, siRNAs may off-target mRNAs and generate deleterious side effects.

Several programs based on empirical data indicate the putative optimal shRNA sequences that are able to generate siRNA strands complementary to the mRNAs. A very important point is to choose a target region of the mRNA, which is not in double-stranded structure and thus accessible to the siRNA. Banks of shRNA genes in lentiviral vectors are available for the mRNAs of different species. It remains that most of the siRNAs do not inhibit the targeted gene to more than 70–80%, which may be insufficient to obtain some relevant animal models. It is tempting to use vectors expressing the shRNA genes at a relatively high level. This may not lead to any significant increase of the inhibition and to an off-targeting, which may be detrimental or even lethal for the animals [89]. In fact, it seems that a well-targeted siRNA can be highly active even at a low concentration. It appears, therefore, to be of paramount importance to select the shRNA capable of inhibiting strongly the targeted mRNA even at a low concentration in cell systems before generating transgenic animals [79].

Use of Transdominant Negative Proteins The action of a gene can be blocked at the protein level by expressing specific inhibitors such as antibodies recognizing the protein of interest [66]. Alternatively, transdominant negative proteins acting as decoys may be used. Transgenic mice mimicking type II diabetes were obtained by overexpressing a mutated insulin receptor still capable of binding the hormone but not of transducing its message [9]. Similarly, overexpression of pseudorabies virus receptor in transgenic mice protects these animals against Aujeszky disease [70].

Genetic Ablation Destroying specifically given cell types in animals may reveal their role in organogenesis. This can be achieved by expressing genes coding for toxins. The challenge is then to express quite specifically the transgenes. Different systems are implemented for this purpose [12, 82]. They rely on two-step mechanisms, which reduce the risk of ectopic expression of the toxin genes.

Control of Transgenes by Exogenous Inducers All the vectors described above and used to express

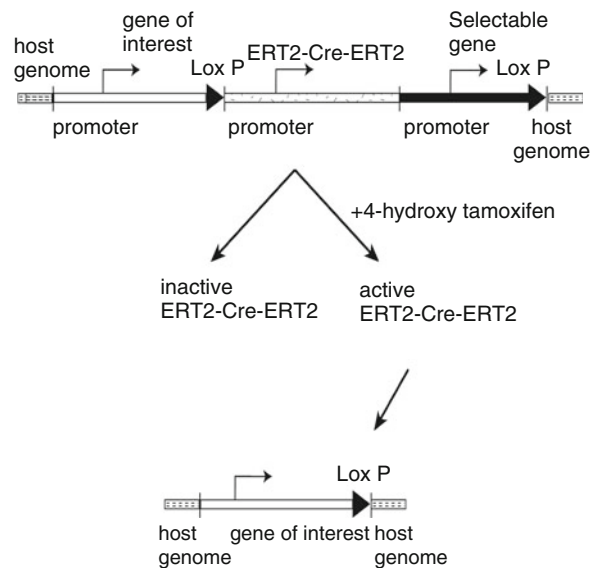


Transgenics: Alternative Gene Transfer Methods. Figure 10

The control of transgene expression may be controlled by an exogenous inducer. In the absence of the inducer (doxycycline), the transcription enhancer (Tet on) is not bound to DNA and it does not stimulate transcription whereas the transcription inhibitor (KRAB) is bound to DNA and reduces the background expression of the gene of interest. In the presence of doxycycline, the reverse is true and the gene of interest is activated

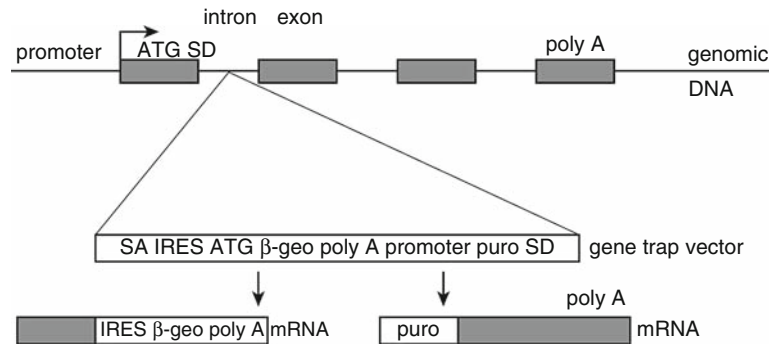
transgenes contain promoters, which are naturally active in the cells of the transgenic animals. This implies that the transgenes are regulated by the natural inducers of the host genes. The induction of a transgene may then coincide with the unwanted stimulation of a number of host genes. Artificial promoters containing regulatory elements from both animal genes and bacterial genes have been designed. The resulting promoters are active in animal cells but controlled by substances active in bacteria but not in animals. The most popular system is based on the use of the bacterial tetracycline repressor gene. In practice, the transgene becomes reversibly activated only when tetracycline or doxycycline is administered to the animals. Various mutants are available making it possible either the induction of the transgene or its inhibition by the addition of the exogenous inducer. A number of similar systems are available and currently used in transgenic animals with a good success [30, 57]. In order to reduce the background expression of the transgene in the absence of the exogenous inducers, the alternative expression of a transcription repressor as KRAB and of an inducer as Tet-on system may be implemented (Fig. 10). These tools, which require the transfer of several genes offer virtually the possibility to express a transgene precisely in a given cell type and at a given moment.

Gene Deletion Conventional homologous recombination makes it possible gene deletion known as knock-out. Another possibility consists of using the Cre-LoxP



Transgenics: Alternative Gene Transfer Methods. Figure 11

Activation of Cre recombinase and selectable gene elimination by 4-hydroxy tamoxifen. The Cre recombinase gene expression may be under the control of the Tet-on system itself under the control of a cell-specific promoter. The fusion protein Cre recombinase–mutated estrogen receptor is active only in the presence of 4-hydroxy tamoxifen. The elimination of the DNA region bordered by LoxP sequences is thus sharply controlled. The selection gene and the Cre recombinase may thus be eliminated from transgenic animals at any stage of their life



Transgenics: Alternative Gene Transfer Methods. Figure 12

Schematic representation of a gene trapping vector. The vector is randomly integrated into the genome of cells that are able to generate transgenic animals (Fig. 3). The β -geo gene is a fusion of the β -galactosidase gene, giving a blue color to cells, and neomycin resistance gene, giving a resistance to geneticin. The vector contains also the puromycin resistance gene, splicing acceptor sites, no promoter, and no transcription terminator. The β -geo selection gene is expressed only when it was integrated into a functional gene. The inactivated gene may be identified by sequencing the genomic DNA flanking the integrated vector. A correlation between the inactivated gene and an altered biological function of the animal may be established

or Flp-FRT systems. A LoxP sequence must first be added to both ends of the fragment to delete. The presence of the Cre recombinase will then recombine the two LoxP sites, leading to a deletion of the DNA fragment located between the LoxP regions. This makes it possible the elimination of a selection gene. The same approach allows the specific and controlled deletion of an inhibitory DNA region, leading to the activation of the gene located in its vicinity. The Cre recombinase may be synthesized by the corresponding gene under the direction of a cell-specific promoter including promoters under the control of doxycycline. Another level of control can be obtained by using an engineered Cre recombinase, which becomes reversibly active in the presence of an estrogen analogue, 4-hydroxy tamoxifen (Fig. 11) [62]. This offers the advantage of having the active Cre recombinase for short periods of time. This prevents the non-specific action of the Cre recombinase, which can recognize cryptic sites in the host genome and induce illegitimate recombination damaging the host DNA. This tool is appropriate to delete genes for resistance to antibiotics.

Vectors for Gene Trapping The identification of genes involved in a given biological function is an essential step to understand the role of the genes

particularly those responsible for human diseases. One possibility is to use vectors for the trapping of active genes. One such vector is described in Fig. 12. Other vectors are shown in the chapter of a book [36]. New tools make it possible random gene knockout in genomes by inserting foreign DNA sequences using transposons [22] or lentiviral vectors. Banks of siRNA genes in lentiviral vectors can also be used to knock-down genes. The inhibited genes can thus be identified and correlations with altered biological functions in the animals become possible.

Future Directions

During the coming decade, transgenesis in animals is expected to be as intensively or even more used than presently. One of the trends is to refine the different tools for basic research as for the different applications. The time when transgenic mice were prepared in a blind manner by microinjecting simple gene constructions is getting over. The recent and very significant improvement of the gene transfer techniques should make several animal species less marginal in the transgenesis field. The number of transgenic models for basic research is being extended to a larger number of species, namely to rats, chicken, Xenopus,

and even cow [47, 80, 96]. The sequencing of an increasing number of genomes, including different mouse strains and even different individuals, will contribute to extend the use of animal transgenesis.

The proportion of transgenic animals used for basic research, including the models for the study of human diseases, is currently very high. This proportion should change but moderately in favor of applications in the food domain and also in the pharmaceutical field including the production of pig cells and organs for patients and the production of recombinant pharmaceutical proteins. Indeed, the production of pig cells and organs is making significant progress but it remains unpredictable when and if the transplantation of pig cells and organs will become a tangible reality as many problems remain to be solved [73].

The production of pharmaceutical proteins in milk and egg white has become a realistic approach as the techniques have reached sufficient maturity even if the production of each protein is a challenge. Yet, this application of transgenesis is coming more slowly than anticipated. For complex reasons not clearly related to the efficiency of the technique or to biosafety guidelines, the pharmaceutical companies remain more or less reluctant to start using animals to obtain pharmaceutical proteins. Companies may consider that their profit is presently higher with the production of recombinant proteins in conventional fermentors harboring animal cell lines, even or because this system generates a shortage of medicaments for patients. A recent multiauthor book is a survey of all the aspects of this technique including the ethical considerations [77].

Whatever happens, the application of transgenic animals for food production should become a reality in the coming decade. The increasing knowledge of alleles in farm animals and of a variety of other possible transgenes should give more space to the use of transgenesis to improve animal production. The recent achievement of the cow genome sequencing will contribute to using cattle alleles more extensively. It remains that the generation of relevant transgenic founders and their extensive use in breeding will be slower than in plants due to the high cost of transgenesis and the time required for the dissemination of their genomes in herds. The poor acceptability of biotechnology particularly in the EU is another hurdle.

The recent progress of several techniques of gene transfer are becoming more popular in laboratories and companies and therefore should be pursued. DNA microinjection into embryos will still be used in the species in which it is efficient. Lentiviral vectors are more and more frequently used, and this trend should become still stronger as the preparation of efficient and safe lentiviral particles has become a standard technique. Other viral vectors could be implemented in future. One candidate is AAV (adeno-associated virus). This virus is spontaneously integrated into cell genomes as lentiviruses. AAV vectors are used with some success for gene therapy in humans [111]. One of its advantages over lentiviruses is that they can harbor longer DNA fragments. Viral vectors might prove attractive for some applications in breeding. It is indeed conceivable to introduce massively foreign genes only into somatic cells. Examples are fish or more generally the animals, which swim or fly and can disseminate in environment. Growth hormone genes transferred into a sufficient number of salmon somatic cells might have the same biological effect on growth than the corresponding transgene without any risk of its dissemination in oceans.

More and more transposons are found and engineered to be able to generate efficiently and in a safe manner a variety of transgenic animals. One of the challenges of modern biology is to decipher the mechanisms involved between given genetic information and the expression of a physiological function. The systematic random integration of transposons and also of lentiviral vectors into the genome of cells able to participate in the development of transgenic animals is expected to generate banks of cells and animals in which physiological disorders can be connected to the genes interrupted by the vectors. This approach is similar to that of generating banks of animals having genes inactivated by knockout or knockdown. Several animal and data banks are available: EMMA in EU, Jackson Laboratory in USA, IMSR (International Mouse Strain Resource), ISTT (International Society for Transgenic Technologies) www.transtechsociety.org, [40], private companies (namely Charles River) as well as academic and private structures for the generation of transgenic mice, rats, and rabbits.

Gene transfer using ICSI has met a great success in mice and a few other species. It is simpler and more

efficient than conventional DNA injection. A broader implementation of this technique is presently limited by the fact that ICSI independently of transgenesis is still possible in only a small number of species. This situation is essentially due to the fact that researchers have not yet acquired this knowledge. The principle of ICSI is common to all species, but the manipulation of oocytes and sperm needs specific training. ICSI should thus be used more intensively in the near future in mice and extended to other species. The possible combination of ICSI and transposons as well as of envelope-free lentiviral particles might contribute to enhance gene transfer efficiency.

A trend is clearly to perform the genetic modifications not directly in embryos but rather in intermediate cells further used to generate transgenic animals. The long-term success of mouse ES cells to target gene transfer via the generation of chimeric animals is expected to be rapidly extended to the use of rat ES cells and iPS cells in different species, which are able to generate chimeric animals as ES cells. Progress must be made before iPS cells are used for transgenesis. This is being achieved as iPS cells are extensively studied since they are expected to be a major tool for cell therapy and gene therapy in humans. The techniques to generate iPS cells from somatic cells must be adapted to each species. An important recent progress is the possibility of dedifferentiating somatic cells no more by transferring the three identified key genes but by introducing the corresponding mRNAs, which are unstable by essence and thus leave no genetic information in the iPS cells. It is conceivable to transfer the three genes able to generate iPS cells into embryonic cells to obtain more easily ES cells. Indeed, cultured ES cells are known to quickly lose their pluripotency in most mouse lines and essentially in all other species but rats. ES cells from a number of species might become a reality in this way. The preliminary success, which made it possible the use of small molecules in culture medium instead of proteins to generate pluripotent cells from embryos or somatic cells is very encouraging. It suggests that the generation and the use of pluripotent cells in multiple species might become a relatively easy task. This would facilitate markedly the genetic modification of animal genomes. A more speculative reasoning is to postulate that the complete dedifferentiation of somatic cells into totipotent cells will be

possible some day by the transfer of genes or chemical inducers, which remain to be found. Cloning and transgenesis as well as their coupling and related ethical problems would be considerably simplified.

Cloning by SCNT proved that it is possible to make the random and targeted gene transfer. The popularization of this technique is likely, but it depends on improvement of the technique. Indeed, clones are often if not always epigenetically modified leading to a limited efficiency, to a reduced welfare of the animals, and potentially to health problems. Significant improvement has been achieved in 10 years, and a better control of cloning should occur in the coming years.

The construction of vectors making it possible a reliable and well-controlled expression of the transgenes will be progressively improved by the empirical use of BAC containing long genomic DNA fragments with the majority of the remote transcription regulators. The use of long genomic DNA fragments as vectors is expected to become more and more frequent but to remain empirical for some time. Most likely, an increasing number of BAC vectors containing regulatory elements for the expression of transgenes in the different cells types will become available. Despite the higher difficulty to manipulate BAC rather than plasmids constructs, researchers will prefer in a number of cases to invest a part of their time in the construction of BAC vectors to tentatively obtain better transgenic models. Rapid progress is being made on the description of chromatin structure and activity. Gene study is more and more performed at the level of a locus and not only of a given gene. This will provide researchers with well-identified chromatin regulators, which will be used to design compact vectors containing these regulators capable of optimizing transgene expression.

Gene targeting is essential to obtain relevant transgenic models. The demonstration that a local cleavage of DNA by natural or engineered meganucleases or by engineered zinc finger nuclease considerably increases the efficiency of homologous recombination is of major importance for basic and applied research particularly for projects involving transgenesis. Gene knockin and knockout are expected to become much more efficient, up to the point perhaps to be possible directly in one-cell embryos and not only via intermediate cells.

The demonstration that DNA cleavage by zinc finger nuclease in the absence of recombination vectors is followed by an imperfect DNA repair (NHEJ), which in a significant proportion of cases leads to a mutation, thus to a gene knockout. The fact that gene knockout by NHEJ could be obtained directly in fish embryos opens new avenues as this protocol does not require any more the laborious gene knockout by homologous recombination. Hundreds of engineered zinc fingers have been designed to recognize specific sites in different genomes. This number should increase rapidly in the coming years. A particularly fascinating point is that NHEJ as well as conventional knockout and knockin might be applied not only to genes proper but also to their genomic regulators. This would allow the study of the regulators of various loci without implementing the laborious manipulation of BACs.

The discovery of siRNA has rapidly been followed by the generation of gene knockdown in plants. The same did not occur in animals. The use of long double-stranded RNA induces deleterious side effects in animals. The availability of lentiviral vector banks harboring genes coding for shRNAs makes gene knockdown easier in animals. The efficiency of this approach remains limited by the insufficient knowledge on the mechanisms of RNA interference. Progress has been made to empirically design siRNA having a high and specific inhibitory effect on gene expression. An important parameter is now taken into consideration. It is known that the sequence of the siRNA determines the choice of the strand, which will target the mRNA of interest and its capacity to inactivate this mRNA. It is now clear that the local structure of the mRNA is essential to make the access of the siRNA possible. Additional progress is expected in this field, and it should facilitate gene knockdown in transgenic animals.

Using constructs in which the transgene is under the control of exogenous inducers like doxycycline is a reality. Sophisticated systems reducing the background expression of the transgene expression in the absence of the inducers are currently used. The possibility of using these systems to create relevant transgenic models is expected to be more and more frequently used. Improvement of these techniques is likely by finding additional inducers.

It is essential to be able to knock out genes in chosen cells and at defined period of the animal life.

This conditional knockout implies the use of the Cre/LoxP or similar systems. The conditional expression of the Cre recombinase and its conditional activation by 4-hydroxy tamoxifen offer a great flexibility. Improvement of these systems is possible. A more extensive use of these systems, which give satisfaction to researchers will be done in the coming years.

Transgenesis techniques are thus making important progress for the generation of transgenic animals as for the fine control of transgene expression. The available tools and those in development are expected to be adapted to the systematic study of multiple genes required for the development of integrative biology.

Bibliography

Primary Literature

1. Baer A, Bode J (2001) Coping with kinetic and thermodynamic barriers: RMCE, an efficient strategy for the targeted integration of transgenes. *Curr Opin Biotechnol* 12:473–448
2. Baker M (2009) Authentic rat embryonic stem cells. *Nat Rep Stem Cells*. doi:10.1038/stemcells.2009.11
3. Bishop JO (1997) Chromosomal insertion of foreign DNA. In: Houdebine LM (ed) *Transgenic animal generation and use*. Harwood Academic Publishers, Amsterdam, pp 219–224
4. Bosselman RA, Hsu RY, Boggs T, Hu S, Bruszewski J, Ou S, Kozar L, Martin F, Green C, Jacobsen F, Nicolson M, Schultz JA, Semon KM, Rishell W, Stewart RG (1989) Germline transmission of exogenous genes in the chicken. *Science* 243:533–535
5. Bronson SK, Smithies O (1994) Altering mice by homologous recombination using embryonic stem cells. *J Biol Chem* 269:27155–27158
6. Bronson SK, Plaehn EG, Kluckman KD, Hagaman JR, Maeda N, Smithies O (1996) Single-copy transgenic mice with chosen-site integration. *Proc Natl Acad Sci USA* 93:9067–9072
7. Buehr M, Meek S, Blair K, Yang J, Ure J, Silva J, McLay R, Hall J, Ying QL, Smith A (2008) Capture of authentic embryonic stem cells from rat blastocysts. *Cell* 135:1287–1298
8. Capecchi MR (1989) Altering the genome by homologous recombination. *Science* 244:1288–1292
9. Chang PY, Benecke H, Le Marchand-Brustel Y, Lawitts J, Moller DE (1994) Expression of a dominant-negative mutant human insulin receptor in the muscle of transgenic mice. *J Biol Chem* 269:16034–16040
10. Chang YF, Imam JS, Wilkinson M (2007) The nonsense-mediated decay RNA surveillance pathway. *Annu Rev Biochem* 76:51–74
11. Chapman SC, Lawson A, Macarthur WC, Wiese RJ, Loechel RH, Burgos-Trinidad M, Wakefield JK, Ramabhadran R, Mauch TJ, Schoenwolf GC (2005) Ubiquitous GFP expression in transgenic chickens using a lentiviral vector. *Development* 132:935–940

12. Chen YT, Levasseur R, Vaishnav S, Karsenty G, Bradley A (2004) Bigenic Cre/loxP, puDeltat conditional genetic ablation. *Nucleic Acids Res* 32:e161
13. Chkheidze AN, Lyakhov DL, Makeyev AV, Morales J, Kong J, Liebhaber SA (1999) Assembly of the α -globin mRNA stability complex reflects binary interaction between the pyrimidine-rich 3' untranslated region determinant and poly(C) binding protein α CP. *Mol Cell Biol* 19:4572–4581
14. Choulika A, Perrin A, Dujon B, Nicolas JF (1994) The yeast I-Sce I meganuclease induces site-directed chromosomal recombination in mammalian cells. *C R Acad Sci* 317:1013–1019
15. Choulika A, Perrin A, Dujon B, Nicolas JF (1995) Induction of homologous recombination in mammalian cells using the I-Sce I system of *Saccharomyces cerevisiae*. *Mol Cell Biol* 15:1968–1973
16. Cohen-Tannoudji M, Robine S, Choulika A, Babinet C, Louvard D, Jaissier F (1998) I-Sce I-induced gene replacement at a natural locus in embryonic stem cells. *Mol Cell Biol* 18:1444–1448
17. Cohen-Tannoudji M, Babinet C, Morello D (2000) Lac Z and ubiquitously expressed genes: should divorce be pronounced? *Transgenic Res* 9:233–235
18. Couble P (1997) The biolistic. In: Houdebine LM (ed) *Transgenic animal generation and use*. Harwood Academic Publishers, Amsterdam, pp 209–213
19. Dann CT (2007) New technology for an old favorite: lentiviral transgenesis and RNAi in rats. *Transgenic Res* 16:571–580
20. de Laat W, Grosveld F (2003) Spatial organization of gene expression: the active chromatin hub. *Chromosome Res* 11:447–459
21. de Laat W, Klous P, Kooren J, Noordermeer D, Palstra RJ, Simonis M, Splinter E, Grosveld F (2008) Three-dimensional organization of gene expression in erythroid cells. *Curr Top Dev Biol* 82:117–139
22. Ding S, Wu X, Li G, Han M, Zhuang Y, Xu T (2005) Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice. *Cell* 122:473–483
23. Dupuy AJ, Clark K, Carlson CM, Fritz S, Davidson AE, Markley KM, Finley K, Fletcher CF, Ekker SC, Hackett PB, Horn S, Largaespada DA (2002) Mammalian germ-line transgenesis by transposition. *Proc Natl Acad Sci USA* 99:4495–4499
24. Echelard Y (1997) Genetic mosaicism in the generation of transgenic mice. In: Houdebine LM (ed) *Transgenic animal generation and use*. Harwood Academic Publishers, Amsterdam, pp 233–225
25. Epinat JC, Arnould S, Chames P, Rochaix P, Desfontaines D, Puzin C, Patin A, Zanghellini A, Paques F, Lacroix E (2003) A novel engineered meganuclease induces homologous recombination in yeast and mammalian cells. *Nucleic Acids Res* 31:2952–2962
26. Garrick D, Fiering S, Martin DI, Whitelaw E (1998) Repeat-induced gene silencing in mammals. *Nat Genet* 18:56–59
27. Gaszner M, Felsenfeld G (2006) Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat Rev Genet* 7:703–713
28. Geurts AM, Cost GJ, Freyvert Y, Zeitler B, Miller JC, Choi VM, Jenkins SS, Wood A, Cui X, Meng X, Vincent A, Lam S, Michalkiewicz M, Schilling R, Foeckler J, Kalloway S, Weiler H, Ménoret S, Anegón I, Davis GD, Zhang L, Rebar EJ, Gregory PD, Urnov FD, Jacob HJ, Buelow R (2009) Knockout rats via embryo microinjection of zinc-finger nucleases. *Science* 325:433
29. Giraldo P, Rival-Gervier S, Houdebine LM, Montoliu L (2003) The potential benefits of insulators on heterologous constructs in transgenic. *Transgenic Res* 12:751–755
30. Gitzinger M, Kemmer C, El-Baba MD, Weber W, Fussenegger M (2009) Controlling transgene expression in subcutaneous implants using a skin lotion containing the apple metabolite phloretin. *Proc Natl Acad Sci USA* 106:10638–10643
31. Gordon JW, Scangos GA, Plotkin DJ et al (1980) Genetic transformation of mouse embryos by microinjection of purified DNA. *Proc Natl Acad Sci USA* 77:7380–7384
32. Hammer RE, Pursel VG, Rexroad CE et al (1985) Production of transgenic rabbits, sheep and pigs by microinjection. *Nature* 315:680–683
33. Han JY (2009) Germ cells and transgenic chicken. *Comp Immunol Microbiol Infect Dis* 32:61–80
34. Henry I, Forlani S, Vaillant S, Muschler J, Choulika A, Nicolas JF (1999) LagoZ and LagZ, 2 genes depleted of CpG dinucleotides, derived from the LacZ gene for the study of epigenetic control. *C R Acad Sci III* 322:1061–1070
35. Herold MJ, van den Brandt J, Seibler J, Reichardt HM (2008) Inducible and reversible gene silencing by stable integration of an shRNA-encoding lentivirus in transgenic rats. *Proc Natl Acad Sci USA* 105:18507–18512
36. Houdebine LM (2003) *Animal transgenesis and cloning*. Wiley, Chichester, 250 pp
37. Houdebine LM (2007) Transgenic animal models and target validation. *Methods Mol Biol* 360:163–202
38. Houdebine LM (2009) Methods to generate transgenic animals in ethics of science and technology assessment. Springer-Verlag, Heidelberg/Berlin/New York, pp 31–48
39. Houdebine LM, Attal J (1999) Internal ribosome entry sites (IRESs): reality and use. *Transgenic Res* 8:157–177
40. Jerchow B (2009) Transgenic technology meeting 2008 in Toronto, Canada: a meeting report. *Transgenic Res*. doi:10.1007/s11248-008-9238-8
41. Kalina J, Senigl F, Micáková A, Mucksová J, Blazkova J, Yan H, Poplstein M, Hejnar J, Trefil P (2007) Retrovirus-mediated in vitro gene transfer into chicken male germ line cells. *Reproduction* 134:445–453
42. Kapuscinski AR, Hayes KR, Li S, Dana G (2007) Environmental risk assessment of genetically modified organisms, vol. 3 Methodologies for transgenic fish. Series eds. Hallerman EM (Guest) and Schei PJ. CAB International, Oxford, pp 304
43. Kayser K (1997) Gene transfer in *Drosophila melanogaster*. In: Houdebine LM (ed) *Transgenic animal generation and use*. Harwood Academic Publishers, Amsterdam, pp 133–137
44. Kuroiwa Y, Kasinathan P, Choi YJ, Naeem R, Tomizuka K, Sullivan EJ, Knott JG, Duteau A, Goldsby RA, Osborne BA, Ishida I, Robl JM (2002) Cloned transchromosomal calves producing human immunoglobulin. *Nat Biotechnol* 20: 889–894

45. Lavitrano M, Bacci ML, Forni M, Lazzereschi D, Di Stefano C, Fioretti D, Giancotti P, Marfe G, Pucci L, Renzi L, Wang H, Stoppacciaro A, Stassi G, Sargiacomo M, Sinibaldi P, Turchi V, Giovannoni R, Della Casa G, Seren E, Rossi G (2002) Efficient production by sperm-mediated gene transfer of human decay accelerating factor (hDAF) transgenic pigs for xenotransplantation. *Proc Natl Acad Sci USA* 99:14230–14235
46. Lee YM, Jung JG, Kim JN, Park TS, Kim TM, Shin SS, Kang DK, Lim JM, Han JY (2006) A testis-mediated germline chimera production based on transfer of chicken testicular cells directly into heterologous testes. *Biol Reprod* 75:380–386
47. Lewin HA (2009) It's a Bull's Market. Research on domesticated livestock heralds new advances in evolutionary biology, animal breeding, and animal models for human diseases. *Science* 324:478–479
48. Li BC, Sun GB, Sun HC, Xu Q, Gao B, Zhou GY, Zhao WM, Wu XS, Bao WB, Yu F, Wang KH, Chen GH (2007) Efficient generation of transgenic chickens using the spermatogonial stem cells in vivo and ex vivo transfection. *Reproduction* 134: 445–453
49. Li P, Tong C, Mehrian-Shai R, Jia L, Wu N, Yan Y, Maxson RE, Schulze EN, Song H, Hsieh CL, Pera MF, Ying QL (2008) Germline competent embryonic stem cells derived from rat blastocysts. *Cell* 135:1299–1310
50. Li W, Wei W, Zhu S, Zhu J, Shi Y, Lin T, Hao E, Hayek A, Deng H, Ding S (2009) Generation of rat and human induced pluripotent stem cells by combining genetic reprogramming and chemical inhibitors. *Cell Stem Cell* 4:16–19
51. Liao J, Cui C, Chen S, Ren J, Chen J, Gao Y, Li H, Jia N, Cheng L, Xiao H, Xiao L (2008) Generation of induced pluripotent stem cell lines from adult rat cells. *Cell Stem Cell* 4:11–15
52. Lillico SG, Sherman A, McGrew MJ, Robertson CD, Smith J, Haslam C, Barnard P, Radcliffe PA, Mitrophanous A, Elliot EA, Sang HM (2007) Oviduct-specific expression of two therapeutic proteins in transgenic hens. *Proc Natl Acad Sci USA* 104:1771–1776
53. Long X, Miano JM (2007) Remote control of gene expression. *J Biol Chem* 282:15941–15945
54. Lunyak VV, Prefontaine GG, Núñez E, Cramer T, Ju BG, Ohgi KA, Hutt K, Roy R, García-Díaz A, Zhu X, Yung Y, Montoliu L, Glass CK, Rosenfeld MG (2007) Developmentally regulated activation of a SINE B2 repeat as a domain boundary in organogenesis. *Science* 317:248–251
55. Lyssiotis CA, Foreman RK, Staerk J, Garcia M, Mathur D, Markoulaki S, Hanna J, Lairson LL, Charette BD, Bouchez LC, Bollong M, Kunick C, Brinker A, Cho CY, Schultz PG, Jaenisch R (2009) Reprogramming of murine fibroblasts to induced pluripotent stem cells with chemical complementation of Klf4. *Proc Natl Acad Sci USA* 106:8912–8917
56. Maclean N (2003) Genetically modified fish and their effects on food quality and human health and nutrition. *Trends Food Sci Technol* 14:242–252
57. Malphettes L, Fussenegger M (2006) Improved transgene expression fine-tuning in mammalian cells using a novel transcription-translation network. *J Biotechnol* 124:732–746
58. Manzini S, Vargiolu A, Stehle IM, Bacci ML, Cerrito MG, Giovannoni R, Zannoni A, Bianco MR, Forni M, Donini P, Papa M, Lipps HJ, Lavitrano M (2006) Genetically modified pigs produced with a nonviral episomal vector. *Proc Natl Acad Sci USA* 103:17672–17677
59. Marsh-Armstrong N, Huang H, Berry DL, Brown DD (1999) Germ-line transmission of transgenes in *Xenopus laevis*. *Proc Natl Acad Sci USA* 96:14389–14393
60. McGrew MJ, Sherman A, Ellard FM, Lillico SG, Gilhooley HJ, Kingsman AJ, Mitrophanous KA, Sang H (2004) Efficient production of transgenic chickens using lentiviral vectors. *EMBO Rep* 5:728–733
61. Mersch B, Gepperth A, Suhai Sand Hotz-Wagenblatt A (2008) Automatic detection of exonic splicing enhancers (ESEs) using SVMs. *BMC Bioinform* 9:369. doi:10.1186/1471-2105-9-369
62. Metzger D, Chambon P (2001) Site- and time-specific gene targeting in the mouse. *Methods* 24:71–80
63. Mialhe E, Boulo V, Cadoret JP, Cedeno V, Rousseau C, Motte E, Gendreau S, Bachère E (1997) Gene transfer technology in marine invertebrates. In: Houdebine LM (ed) *Transgenic animal generation and use*. Harwood Academic Publishers, Amsterdam, pp 151–155
64. Moisyadi S, Kaminski JM, Yanagimachi R (2009) Use of intracytoplasmic sperm injection (ICSI) to generate transgenic animals. *Comp Immunol Microbiol Infect Dis* 32:47–60
65. Moreira PN, Pozueta J, Pérez-Crespo M, Valdivieso F, Gutiérrez-Adán A, Montoliu L (2007) Improving the generation of genomic-type transgenic mice by ICSI. *Transgenic Res* 16:163–168
66. Müller M (2000) Increasing disease resistance in transgenic domestic. In: Toutant JP, Balazs E (eds) *Molecular farming*. Institute National de la Recherche Agronomique, Paris, pp 87–98
67. Murphey P, Yamazaki Y, McMahan CA, Walter CA, Yanagimachi R, McCarrey JR (2009) Epigenetic regulation of genetic integrity is reprogrammed during cloning. *Proc Natl Acad Sci USA* 106:4731–4735
68. Nakagawa M, Koyanagi M, Tanabe K, Takahashi K, Ichisaka T, Aoi T, Okita K, Mochiduki Y, Takizawa N, Yamanaka S (2008) Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts. *Nat Biotechnol* 26: 101–110
69. Notini AJ, Li R, Western PS, Sinclair AH, White SJ (2009) Rapid and reliable determination of transgene zygosity in mice by multiplex ligation-dependent probe amplification. *Transgenic Res* 18:987–991
70. Ono E, Amagai K, Taharaguchi S, Tomioka Y, Yoshino S, Watanabe Y, Cherel P, Houdebine LM, Adam M, Eloit M, Inobe M, Uede T (2004) Transgenic mice expressing a soluble form of porcine nectin-1/herpesvirus entry mediator C as a model for pseudorabies-resistant livestock. *Proc Natl Acad Sci USA* 101:16150–16155
71. Palmiter RD, Brinster RE, Trumbauer ME, Rosenfeld MG, Birnberg NC, Evans RM (1982) Dramatic growth of mice that develop from eggs microinjected with metalloproteinase-growth hormone fusion genes. *Nature* 300:611–615

72. Pera MF, Hasegawa K (2008) Simpler and safer cell reprogramming. *Nat Biotechnol* 26:59–60
73. Petersen B, Carnwath W, Niemann H (2009) The perspectives for porcine-to-human xenografts. *Comp Immunol Microbiol Infect Dis* 32:61–80
74. Pfeifer A (2006) Lentiviral transgenesis – a versatile tool for basic research and gene therapy. *Curr Gene Ther* 6: 535–542
75. Porteus MF, Carroll D (2005) Gene targeting using zinc finger nucleases. *Nat Biotechnol* 23:967–973
76. Rao M (2007) Scalable human ES culture for therapeutic use: propagation, differentiation, genetic modification and regulatory issues. *Gene Ther* 15:82–88
77. Reh binder E, Engelhard M, Hagen K, Jorgensen RB, Pardo-Avellaneda R, Schnieke A, Thiele F (2009) Pharming. Promises and risks of biopharmaceuticals derived from genetically modified plants and animals. *Ethics of Science and Technology Assessment Vol 35*. Springer-Verlag, Berlin/Heidelberg
78. Ritchie WA, Neil C, King T, Whitelaw CB (2007) Transgenic embryos and mice produced from low titre lentiviral vectors. *Transgenic Res* 16:661–664
79. RNAi: Multi-author review (2009) *Nature* 457:7228
80. Roberts RM, Smith GW, Bazer FW, Cibelli J, Seidel GE Jr, Bauman DE, Reynolds LP, Ireland JJ (2009) Farm animal research in crisis. *Science* 324:468–469
81. Robl JM, Wang Z, Kasinathan P, Kuroiwa Y (2007) Transgenic animal production and animal biotechnology. *Theriogenology* 67:127–133
82. Saito M, Iwawaki T, Taya C, Yonekawa H, Noda M, Inui Y, Mekada E, Kimata Y, Tsuru A, Kohno K (2001) Diphtheria toxin receptor-mediated conditional and targeted cell ablation in transgenic mice. *Nat Biotechnol* 19:746–750
83. Salter DW, Smith EJ, Hughes SH, Wright SE, Crittenden LB (1987) Transgenic chickens: insertion of retroviral genes into the chicken germ line. *Virology* 157:236–240
84. Santiago Y, Chan E, Liu PQ et al (2008) Targeted gene knockout in mammalian cells by using engineered zinc-finger nucleases. *Proc Natl Acad Sci USA* 105:5809–5814
85. Schnieke AE, Kind AJ, Ritchie WA, Mycock K, Scott AR, Ritchie M, Wilmut I, Colman A, Campbell KH (1997) Human factor IX transgenic sheep produced by transfer of nuclei from transfected fetal fibroblasts. *Science* 278:2130–2133
86. Scott BB, Lois C (2005) Generation of tissue-specific transgenic birds with lentiviral vectors. *Proc Natl Acad Sci USA* 102:16443–16447
87. Shen W, Li L, Pan Q, Min L, Dong H, Deng J (2006) Efficient and simple production of transgenic mice and rabbits using the new DMSO-sperm mediated exogenous DNA transfer method. *Mol Reprod Dev* 73:589–594
88. Shinohara ET, Kaminski JM, Segal DJ, Pelczar P, Kolhe R, Ryan T, Coates CJ, Fraser MJ, Handler AM, Yanagimachi R, Moisyadi S (2007) Active integration: new strategies for transgenesis. *Transgenic Res* 16:333–339
89. Sioud M (2006) Innate sensing of self and non-self RNAs by toll-like receptors. *Trends Mol Med* 12:167–176
90. Sippel AE, Saueressig H, Hubler MC, Faust N, Bonifer C (1997) Insulation of transgenes from chromosomal position effects. In: Houdebine LM (ed) *Transgenic animal generation and use*. Harwood Academic Publishers, Amsterdam, pp 257–266
91. Smith K, Spadafora C (2005) Sperm mediated gene transfer: applications and implications. *BioEssays* 27:551–562
92. Taboit-Dameron F, Malassagne B, Viglietta C, Puissant C, Leroux-Coyau M, Chéreau C, Attal J, Weill B, Houdebine LM (1999) Association of the 5'HS4 sequence of the chicken beta-globin locus control region with human EF1-alpha gene promoter induces ubiquitous and high expression of human CD55 and CD59 cDNAs in transgenic rabbits. *Transgenic Res* 8:223–235
93. Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, Yamanaka S (2007) Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* 131:861–872
94. Takehashi M, Kanatsu-Shinohara M, Inoue K, Ogonuki N, Miki H, Toyokuni S, Ogura A, Shinohara T (2007) Adenovirus-mediated gene delivery into mouse spermatogonial stem cells. *Proc Natl Acad Sci USA* 104:2596–2601
95. Tamura T, Thibert C, Royer C, Kanda T, Abraham E, Kamba M, Komoto N, Thomas JL, Mauchamp B, Chavancy G, Shirk P, Fraser M, Prudhomme JC, Couble P (2000) Germline transformation of the silkworm *Bombyx mori* L. using a piggyBac transposon-derived vector. *Nat Biotechnol* 1:81–84
96. The Bovine Genome Sequencing and Analysis Consortium, Elsik CG, Tellam RL, Worley KC (2009) The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* 324:522–528
97. Thermes V, Grabher C, Ristatore F, Bourrat F, Chouluka A, Wittbrodt J, Joly JS (2002) I-SceI meganuclease mediates highly efficient transgenesis in fish. *Mech Dev* 118:91–98
98. Thierry-Mieg D, Naert K, Bonnerot C (1997) Genetic transformation of *Caenorhabditis elegans*. In: Houdebine LM (ed) *Transgenic animal generation and use*. Harwood Academic Publishers, Amsterdam, pp 137–151
99. Tiscornia G, Singer O, Ikawa M, Verma IM (2003) A general method for gene knockdown in mice by using lentiviral vectors expressing small interfering RNA. *Proc Natl Acad Sci USA* 100:1844–1848
100. Umland T, Montoliu L, Schütz G (1997) The use of yeast artificial chromosomes for transgenesis. In: Houdebine LM (ed) *Transgenic animal generation and use*. Harwood Academic Publishers, Amsterdam, pp 289–298
101. Van de Lavoie MC, Diamond JH, Leighton PA, Mather-Love C, Heyer BS, Bradshaw R, Kerchner A, Hooi LT, Gessara TM, Swanberg SE, Delany ME, Etches RJ (2006) Germline transmission of genetically modified primordial germ cells. *Nature* 441:766–769
102. Van Keuren ML, Gavriliina GB, Filipiak WE, Zeidler MG, Saunders TL (2009) Generating transgenic mice from bacterial artificial chromosomes: transgenesis efficiency, integration and expression outcomes. *Transgenic Res*. doi 10.1007/s11248-009-9271-2
103. Van Reenen CG (2009) Assessing the welfare of transgenic farm animals. In *Ethics of science and technology*

assessment. Springer-Verlag, Heidelberg/Berlin/New York, pp 119–143

104. Van Reenen CG, Meuwissen THE, Hopster H, Oldenbroek K, Kruip TH, Blokhuis HJ (2001) Transgenesis may affect farm animal welfare: a case for systemic risk assessment. *J Anim Sci* 79:1763–1769
105. Wernig M, Meissner A, Foreman R, Brambrink T, Ku M, Hochedlinger K, Bernstein BE, Jaenisch R (2007) In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. *Nature* 448:318–324
106. Wilson JH (2008) Knockout punches with a fistful of zinc fingers. *Proc Natl Acad Sci USA* 105:5653–5654
107. Woods IG, Schier AF (2008) Targeted mutagenesis in zebrafish. *Nat Biotechnol* 26:650–651
108. Yang SH, Agca Y, Cheng PH, Yang JJ, Agca C, Wing Sang Chan A (2007) Enhanced transgenesis by intracytoplasmic injection of envelope-free lentivirus. *Genesis* 45:177–183
109. Yong HY, Hao Y, Lai L, Li R, Murphy CN, Rieke A, Wax D, Samuel M, Prather RS (2006) Production of a transgenic piglet by a sperm injection technique in which no chemical or physical treatments were used for oocytes or sperm. *Mol Reprod Dev* 73:595–599
110. Zambrowicz BP, Imamoto A, Fiering S, Herzenberg LA, Kerr WG, Soriano P (1997) Disruption of overlapping transcripts in the ROSA beta geo 26 gene trap strain leads to widespread expression of beta-galactosidase in mouse embryos and hematopoietic cells. *Proc Natl Acad Sci USA* 94:3789–3794

Books and Reviews

- Honaramooz A, Megee S, Zeng W, Destrempe MM, Overton SA, Luo J, Galantino-Homer H, Modelski M, Chen F, Blash S, Melican DT, Gavin WG, Ayres S, Yang F, Wang PJ, Echeland Y, Dobrinski I (2008) Adeno-associated virus (AAV)-mediated transduction of male germ line stem cells results in transgene transmission after germ cell transplantation. *FASEB J* 22:374–382
- Niemann H, Kues W, Carnwath JW (2009) Methods to generate transgenic animals in ethics of science and technology assessment, vol. 34. Springer-Verlag, Heidelberg/Berlin/New York, pp 1–30

Transit-Oriented Development and Land Use

ROBERT CERVERO

Department of City and Regional Planning,
University of California, Berkeley, CA, USA

Article Outline

Glossary

Definition of the Subject

Introduction

The Scope of TOD

Bus-Based TOD: Bogotá and Curitiba

America's TOD Success Story: Arlington County,
Virginia

The Transportation and Environmental Benefits of
TOD

TOD and Ridership at the Workplace

TOD and Parking

TOD Implementation Tools

Central-City Redevelopment Around Streetcar Lines

The Challenges of Mixed-Use TODs

Financing TODs

Future Directions

Bibliography

Glossary

Bus rapid transit High-quality bus services that mimic some of the mobility benefits of exclusive-guideway rail transit services but at a fraction of the cost, including busways, dedicated-lane services, and buses that operate with special features like signal prioritization and queue-jumper lanes at busy intersections.

Joint development Development activities on or near a transit agency's property that generates revenue and ridership benefits.

Public transportation Common-carrier forms of mass transportation available as a mobility service to the general populous, including heavy rail, light rail, and bus transit.

Self selection Willful decisions of residents to move into particular neighborhoods for lifestyle reasons and thereby have a bearing on travel behavior.

Smart growth Compact, mixed-use urban development that encourages alternatives to automobile travel, including transit usage, and that promotes environmental conservation and which can include, but is not limited to, transit-oriented development.

Transit value capture Securing profits from increases in land values conferred by public transit investments to help finance capital costs, including fixed guideways, rolling stock, and neighborhood improvements.

Definition of the Subject

Transit-oriented development, or TOD, is a term used to describe the physical integration and linkage of

public transportation investments and urban land development at or near a station. TOD typically takes the form of compact, mixed land-use development with high-quality pedestrian environments that is concentrated around rail transit nodes [1]. TODs can also occur around the stations and stops of busways and bus services that operate on exclusive, dedicated lanes, often referred to as Bus Rapid Transit (BRT).

The chief objective of TOD is to encourage larger shares of motorized trips to be taken by public transportation and as a result to reduce traffic congestion and improve environmental conditions, including air quality (since the private automobile is the primary source of ozone, carbon monoxide, and particulate matter emissions in many cities of the world). Secondary objectives of TOD might include revitalizing a long-distressed urban district, encouraging the production of more affordable housing and curbing the consumption of agricultural land and open space.

Introduction

TOD is often viewed as an antidote to car-dependent sprawl. By attracting a mix of residences, businesses, shops, and civic activities within a quarter-mile walking distance of an urban railway station, proponents argue that TOD can draw people to transit and thus relieve traffic congestion, improve air quality, and contribute toward climate stabilization. The station and its immediate surroundings can also serve as the hub of a community – a focal point for organizing the regeneration of stagnant neighborhoods and in the case of greenfields, the design and construction of new ones. Thus TOD is championed for both its environmental and its place-making benefits. Of course, the two are not unrelated. Traffic-snarled districts hardly make attractive, livable places. And as discussed later, places that are conducive to transit riding can help free-up traffic jams.

TOD has gained popularity in the United States, the world's most car-dependent country, for three key reasons. One, it is arguably the most cogent, understandable form of "smart growth." Citizens, politicians, and city-planners alike understand that if there is a logical place to concentrate urban growth, it is in and around transit stations. Second, demographic and lifestyle trends are working in favor of TOD. Living around

transit stations appeals to growing numbers of Americans, like childless couples, Generation X'ers, and empty-nesters who value convenience and access to high-quality transit and who place a premium on being in a walkable community with urban amenities. Recent studies have rated TOD as a top real estate investment, estimating that as many as one-third of newly formed households in rail-served US metropolitan areas are receptive to transit-oriented living [2]. For many, TOD residency is a quality-of-life issue. As one study put it, "TODs are all about creating choice for people, about housing, lifestyle and travel"[3]. Third, more and more public policy initiatives favor TODs. As the federal, state, and local government levels, funding programs and incentives in the form of regulatory relief are favoring public and private invests in and around transit stations.

The Scope of TOD

A national study in the United States found that in 2003 about 100 of the nation's 3,300 fixed rail transit stations, or around 3%, could be considered TODs – in the sense all had a mix of land uses, embodied urban designs that created safe and attractive walking environments, and met minimum density thresholds (at least 12 dwelling units and 50 workers per net acre of land) [4]. Far more common than TODs, however, are what have been called "Transit adjacent developments," or TADs. TADs comprise buildings that are near transit stops but fail to induce residents, workers, and shoppers to patronize transit because of their designs and physical orientations. A mid-rise office building with standard parking supplies whose main entrance faces away from the nearby train station, forcing patrons to take a round-about path, is more TAD than TOD. By one definition, TAD is "physically near transit (but) fails to capitalize upon this proximity. . . (It) lacks any functional connectivity to transit – whether in terms of land-use composition, means of station access, or site design" [5].

The trade-off between creating a TOD versus a TAD reflects the inherent conflict between the transit station as "nodes" versus "places." On the one hand, stations are logistical nodes wherein cars, buses, taxis, delivery trucks, pedestrians, and cyclist converge for accessing transit and allowing intermodal transfers. On the other

hand, stations and their environs are places for creating or rebuilding community hubs. Whenever the logistical needs of a station win out, the resulting road designs and parking layouts often detract from the quality of walking, creating more of a TAD than a TOD. This has often been the case in and around suburban rail transit stations in the United States.

On the global stage, TOD is most fully developed in Europe, and in particular Scandinavia. In cities like Copenhagen, Denmark, and Stockholm, Sweden, corridors for channeling overspill growth from the urban centers were identified and reserved early in the planning process, and rail infrastructure was built, often in advance of demand, to steer growth along desired growth axes [6]. As importantly, greenbelt wedges set aside as agricultural preserves, open space, and natural habitats were also designated and accordingly major infrastructure was directed away from these districts. In the case of Stockholm, the last half-century of strategic regional planning has given rise to a regional settlement and commutation pattern that has substantially lowered car-dependency in middle-income suburbs. Stockholm planners focused on creating jobs-housing balance along rail-served axial corridors. This in turn has produced directional-flow balances. During peak hours, 55% of commuters are typically traveling in one direction on trains and 45% are heading in the other direction [6]. Moreover, Stockholm's transit modal share is nearly twice that found in bigger rail-served European cities like Berlin and even higher than inner London's market share. In fact, Stockholm is one of the few places where automobility appears to be receding. Between 1980 and 1990, it was the only city in a sample of 37 global cities that registered a per capita decline in car use – a drop off of 229 annual kilometers of travel per person [7].

Bus-Based TOD: Bogotá and Curitiba

While rail-based TODs are most prevalent worldwide, in good part because rail transit delivers the most regional accessibility benefits and thus are attractive to private investors, in recent time TODs are also taking form along Bus Rapid Transit (BRT) corridors. Many medium-sized global cities are looking to BRT as the most affordable form of high-performance public transit investment. BRT aims to achieve the speed and

performance advantages of grade-separated services at a fraction of the cost by cleverly using bus-based approaches. Among its key features are: exclusivity, notably physical segregation; seamless (same-level) transfers; advanced bus technology: clean fuels, lightweight materials, low floors, advanced communications, docking systems; supportive armature: signal priorities, bus turnouts, curb realignments, automated vehicle location (AVL) systems, automated routing and dispatching; and expeditious fare collection and boarding: off-vehicle payment, smart cards. Two noteworthy international experiences with BOT and TOD, both in Latin America, are Curitiba, Brazil and Bogotá, Colombia.

The environmental benefits of balancing urban growth along bus-served linear axes and aggressively pursuing a “transit first” policy are underscored by experiences in Curitiba, Brazil. Curitiba, widely viewed as one of the world's most sustainable, well-managed metropolises, is also one of the most accessible – a product of some 40 years of carefully integrating urbanization and transportation improvements. By emphasizing planning for people rather than cars, Curitiba has evolved along well-defined radial axes that are intensively served by dedicated busways. Along some corridors, streams of double-articulated buses haul 16,000 passengers per hour, comparable to what much pricier Metrorail systems carry. A design element used to enhance accessibility in Curitiba is the “trinary” – three parallel roadways with compatible land uses. An important benefit of mixed land uses and transit service levels along these corridors, besides phenomenally high ridership rates, has been balanced, bi-directional flows, ensuring efficient use of available bus capacity, just as in the case of Stockholm. On a per capita basis, Curitiba is one of Brazil's wealthiest city yet it averages considerably more transit trips than much bigger Rio de Janeiro and São Paulo [6]. It also boasts the cleanest air among any Brazilian city over 1 million inhabitants, despite being a provincial capital with a sizable industrial sector. The strong, workable nexus that exists between Curitiba's bus-based transit system and its mixed-use linear settlement pattern deserves most of the credit.

Bogotá, the Andean capital of Colombia, has gained global recognition for its highly efficient and productive bus rapid transit (BRT) system. For a city of

7.5 million inhabitants facing civil conflict and deep economic problems, Bogotá's emergence as one of the world's most sustainable metropolises is all the more remarkable. In the late 1990s, Bogotá began operating a high-speed, high-capacity bus system, called Transmilenio, building upon Curitiba's successes with dedicated busways. A big difference, however, is that Curitiba relies principally upon circular, crosstown bus routes to interconnect radial busways. Outside of downtown, relatively little was invested in pedestrian and bikeway improvements. Bogotá, on the other hand, actively embraced pedestrian and bicycle access.

The 42-km, 3-line Transmilenio busway is the centerpiece of Bogotá's vast bus network. (The dedicated system will eventually expand to 22 lines covering 391 km.) Bus lanes are situated in boulevard medians, with weather protected, attractively designed stations spaced every 500 m or so. Because of dual carriageways that enable buses to overtake each other and high-level platforms that allow expeditious boardings and alightings, Transmilenio has a throughput of some 35,000 persons per direction per hour, a number that matches that of many Metrorail system. Some one million passengers ride Transmilenio buses each weekday, three times the ridership of two rail lines in Medellín, Colombia (achieved at less than one-fifth of the Medellín Metro's construction costs) and according to one study providing for a social rate of return of 61% [8].

Particularly important to the transitway has been Bogotá's attention to pedestrian and bicycle access, in the form of "green connectors." Perpendicular and grade-separated pedways and bikeways connect some of the poorest barrios and informal housing settlements (with highly transit-dependent populations) to the busways. Other innovative features of Bogotá's sustainable transport program include license-plate rationing, parking management, and car-free districting. Bogotá is an extraordinary example of matching infrastructure "hardware" with public-policy "software." The city boasts Latin America's most extensive network of cycleways (250 km), the world's longest pedestrian corridor (17 km), and the planet's biggest Car Free Day (covering an entire city of 35,000 ha).

While Bogotá's Transmilenio system has not dramatically altered the cityscape to date, at least when compared to places like Curitiba, research shows that commercial properties have reaped benefits from

proximity to busway stations. A hedonic price-model analysis found a monthly rental discount of 1.87% for every additional 0.1 km from a BRT station, all else being equal [9]. This suggests a pent-up market demand for the accessibility benefits conferred by high-quality bus-based transit in cities of developing countries. A more recent analysis found that land-value capitalization increased as Transmilenio expanded into new neighborhood, revealing the network benefits of adding regional connectivity to the system [10].

As with many successful transit investments, it has been the attention to design details, matched by good macro-scale planning, that has contributed to Transmilenio's success [11]. Car parking is mainly limited to the end stations of the Transmilenio busway. Nearly half of the 57 intermediate stations are served by skywalks/pedestrian overpasses. A phalanx of sidewalks and bikeways feed into all stations, most embellished by vegetative landscaping. Some two dozen civic plazas, pocket parks, and recreational facilities lie within a half kilometer of busway stops. Today, an estimated 45% of Transmilenio users reach stations by foot or bicycle [11].

America's TOD Success Story: Arlington County, Virginia

No place in the United States has witnessed more mid-to-high rise, mixed-use development along a rail corridor over the past three decades than Arlington County, Virginia [4]. This 26 square mile county just south of the nation's capital has experienced a tremendous increase in building activity since Washington Metrorail's 1978 opening: more 25 million square feet of office space, 4+ million square feet of retail space, some 25,000 mixed-income dwelling units, and over 6,500 hotel rooms. Of the nearly 190,000 people today living in Arlington County, in the early 2000s, 26% resided within a Metrorail-served corridor (roughly a one-quarter mile walkshed of stations), even though these corridors comprised only 8% of county land area. If the development added to these two corridors had been built at suburban density standards, such as in neighboring Fairfax County, Virginia, seven times as much land area would have been required [4].

Arlington County's TODs have hardly been the result of good fortune or happenstance.

The transformation of once-rural Arlington County into a showcase of compact, mixed-use TOD has been the product of ambitious, laser-focused station-area planning and investment. Prior to Metrorail's arrival, Arlington County planners understood that high-performance transit provided an unprecedented opportunity to shape future growth and proceeded to introduce various strategies – targeted infrastructure improvements, incentive zoning, development profers, and permissive and as-of-right zoning – to entice private investments around stations. After preparing countywide and station-area plans on desired land-use outcomes, density and setback configurations, and circulation systems, zoning classifications were changed and developments that complied with these classifications could proceed unencumbered. The ability of complying developers to create TODs “as-of-right” was particularly important for it meant developers could line up capital, secure loans, incur upfront costs, and phase-in construction without the fear of local government “changing its mind.” Another key factor was the decision not to align Arlington County's Metrorail rail corridor in the median of Interstate-66, thus suppressing development potential. Instead, County officials persuaded the region's transit authority to align the corridor in the traditional urban centers to help jump-start the process of urban regeneration.

The pay-off of concentrated growth along rail corridors is revealed in Arlington County's transit ridership statistics. The County today boasts one of the highest percentages of transit use in the Washington, DC, region, with nearly 40% of Metrorail corridor residents commuting to work by public transit [4]. An important outcome of promoting mixed-use development along Arlington County's rail corridors has been balanced jobs and housing growth which in turn has produced balanced two-way travel flows. Counts of station entries and exits in Arlington County were nearly equal during peak hours as well as the off-peak. During the morning rush hours, many of the county's Metrorail stations are both trip origins and destinations, meaning trains and buses are full in both directions. The presence of so much retail-entertainment-hotel activities along the County's Metrorail corridors has further filled trains and buses during the midday and on weekends. Arlington County averages higher shares of transit boardings

and alightings at its stations in off-peak hours than other jurisdiction in the region with the exception of downtown Washington, DC.

The Transportation and Environmental Benefits of TOD

The litmus test for whether TOD yields environmental and other societal benefits is the degree to which it increases transit ridership, and in particular draws former motorists into trains and buses. Past research in the United States shows that neighborhoods designed according to TOD principles are associated with lower car ownership rates [4, 12], appreciably higher transit modal splits for commuting [13, 14], and fewer vehicle trips per day [15]. In California, surveys show that residents who live near a transit station use transit for their commutes at a rate four to five times higher than residents of the same region who don't live near stations [14]. This pattern has held steady over time. In the case of the Pleasant Hill station of the Bay Area Rapid Transit (BART) system in the San Francisco-Oakland region, for instance, 47% of station-area residents took transit to work in 1993 [13]. Ten years later, in 2003, the share was 44% [14].

While TODs bump up the shares of residents' trips by transit, more important to many observers – particularly traffic engineers and elected officials – is the impact on local traffic conditions. A recent study of 17 TOD projects in five US metropolitan areas (Philadelphia, Northeast New Jersey, Washington, DC, Portland, and the San Francisco Bay Area) uncovered evidence of “vehicle trip de-generation” [15]. Over a typical weekday period, TOD housing projects in these five regions averaged 44% fewer daily vehicle trips than that estimated by the *Trip Generation* manual of the Institute of Transportation Engineers (ITE), which is widely used as the reference manual for gauging traffic impacts of new projects. During peak hours, transit-oriented housing projects generated one-half the typical number of vehicle trips per dwelling unit.

The higher transit use and lower trip generation rates among station-area residents are largely a product of what economists call “self selection.” People who prefer to take a train to work – whether to avoid the stress of fighting traffic or to have time to read a newspaper en route to work – purposely choose to

live near a rail stop, that is, they are predisposed to transit commuting; they are not “converted” to transit use simply because of where they live. Residing near transit is thus a lifestyle choice. A recent study estimated that around 40% of the ridership bonus attributed to TOD is due to self selection [16].

While TODs generate fewer trips per dwelling unit, this does not necessarily mean they reduce local traffic congestion. Indeed, often the opposite holds which poses a fundamental dilemma – not only for TODs but for all urban developments that increase average densities. Invariably, because not all TOD residents take transit, denser development will congest the nearby road intersections during peak periods. Experiences show that in settings like the United States, where the majority of households own cars, a given percentage increase in density (dwelling units per acre) will not be matched a similar percentage decline in car trips per acre, meaning that car-traffic densities invariably rise. Thus in the near term, even if it is well served by transit, dense development translates into more traffic congestion. Not-in-my-backyard (NIMBY) opposition to higher density has stopped TOD plans around a number of middle-income neighborhoods served by the heavy-rail BART system in the San Francisco East Bay, mainly through building height restrictions. Over the past 20 years, some 2,400 new households have located within a 5-min walk to the Pleasant Hill BART station. Once a critical mass of TOD residents forms, they create neighborhood associations which, among other things, stop efforts to add more development. In Pleasant Hill, plans to build a massive entertainment complex, with a 20-screen IMAX movie theater, were scuttled due to neighborhood backlash. The lesson is clear: do not expect to transform middle-income, stable neighborhoods with large-scale infill projects, because residents have the political might to stop such proposals every step of the way. More acceptable are TOD proposals for transitional neighborhoods, redevelopment districts, or greenfields with few people around to oppose new development. Also, large-scale, regional trip generators that bring outsiders to neighborhoods made up of professional-class residents will rarely be embraced by residentially dominated TODs. Only local- and neighborhood-serving land uses will be welcomed in such settings.

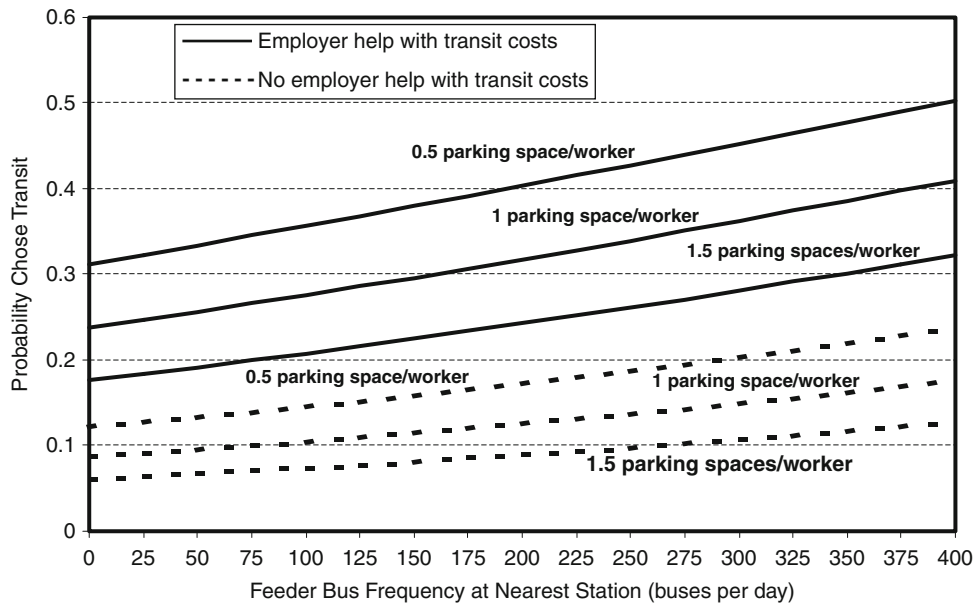
The downside of preventing new growth around transit stations is that the development will end up elsewhere and most likely increase vehicle miles traveled (VMT) – the strongest single correlate of greenhouse gas emissions (GHG), air pollution, and energy consumption in the transport sector. Like toothpaste in a tube, if development is squeezed from one area, it simply gets redistributed elsewhere. Shifting growth from TOD to AOD (automobile-oriented development) can reduce local traffic congestion at the cost of environmental degradation for a region at large.

It can be argued that there is good and bad congestion, just like good and bad cholesterol. Congestion caused by TOD, many urbanologists would contend, is, for the most part, good. There is no avoiding the fact that denser areas have denser traffic. Dense cities with world-class rail systems, like Paris and London, are often more traffic-choked than sprawling cities. Yet these dense cities are also attractive places to live, work, and visit. Congestion is part of the territory of being an active, vibrant place. TOD, moreover, offers a relief valve to congestion. In a TOD, it is easier to predict when and where congestion will occur, and it is easier to avoid that congestion, e.g., one can hop on a train or ride a bike.

TOD and Ridership at the Workplace

While TOD planning and design has traditionally focused on residential development, unless nonresidential destinations are also served by high-quality transit, few TOD residents opt for rail or bus travel. Studies show that TOD workplaces also yield ridership dividends, though not because of self-selection but rather proximity and convenience [17]. Parking policies and designs are particularly important determinants of whether those working in TODs ride transit or not.

A study of workers in 10 office buildings near suburban rail stations in California found, on average, 19% of commute trips were by public transit, with considerable variation [14, 16]. Other factors, however, weighed heavily on whether TOD workers commuted by transit. This is reflected by the logit-model results in Fig. 1. The figures show the likelihood that TOD workers in California commuted by transit as a function of parking supplies, employer assistance with transit costs, and frequency of feeder bus services. With 25 feeder



Transit-Oriented Development and Land Use. Figure 1

Influences of employer policies and feeder bus frequencies on transit commuting among office workers in California

buses per day, an office setting with 50% more parking spaces than workers, and no employer help with transit costs, the model predicts that just 8% of office workers near a rail station will commute by transit. At the other extreme, for a worker heading to a station with 400 daily feeder buses who work for an employer who provides transit-pass assistance and provides one parking space for every two workers, the likelihood that he or she will commute by transit is 50%. Over the range of feeder bus frequencies, the differential in transit commuting probabilities is 30–40% depending on how generous employers are in promoting transit (i.e., minimal parking and help with transit costs) or in accommodating the automobile (i.e., ample parking and no help with transit costs). Clearly, successful TODs are far more than physical design challenges: the policy “software” that accompanies the built-design “hardware” matters tremendously.

TOD and Parking

Critics charge that many large-scale housing projects near urban rail stations are “over-parked” – more parking is provided than is needed [18]. Excessive parking can drive up the cost of housing, consume

valuable land near transit, and impose such environmental costs as increased impervious surface areas (that contribute to urban heat-island effects and stream-water pollution from oily run-offs). As noted, transit stations are thought to offer “location efficiency,” enabling residents to get by with fewer cars than they might otherwise own. Yet lenders and planners often insist that code-standard parking be provided in station areas regardless. This can drive up the cost of station-area housing, thus defeating the purpose of promoting affordable housing construction near major transit stops.

A recent study raises some questions whether suburban TODs indeed average less parking, at least in the United States [15]. In the case of Portland, Oregon, while average vehicle trip generation rates of TOD projects were 41% below ITE rates, the average use of parking spaces was only 11% less. The parking occupancy at three of the 15 surveyed TOD projects in Portland was actually higher than that predicted by the ITE *Parking Generation* manual. In the San Francisco Bay Area, owning and parking a car was even more of a necessity. There, TOD parking rates were equivalent to ITE’s standard of 1.2 spaces per unit. For all seven TOD housing projects surveyed

near the Fremont BART station, parking levels were actually higher than the ITE rates – as much as 40% above.

What might explain high parking rates combined with high transit ridership rates for TODs? One factor may be that most cities do not reduce the parking requirements for TODs which in turn induces car ownership. One survey of TODs in California found no reduction in cities' parking requirements at seven of the 11 sites studied [17]. Planners appear to assume that more transit will not reduce parking demand, and they may be right. In most suburban TODs, many residents still need access to a car. They just do not use them as much. But like most suburbanites, they still need a car to get to most non-work destinations – the vast majority of which are located away from rail stops. While transit-oriented housing might mean that more trip *origins* are near rail stops, as long as most *destinations* are not, TOD residents will still own cars and use them for shopping, going out to eat, and the like.

The fact that some suburban TODs encourage transit riding yet fail to prompt residents to reduce car ownership suggests that TODs might be a natural setting for carsharing. A study of City CarShare members in San Francisco found that 4 years into the program, 29% of carshare members had gotten rid of one or more of their cars and 63% lived in zero-vehicle households [18]. Putting shared-cars in and around TODs could relieve many households from owning a second car or a vehicle altogether. Through a combination of proximity advantages and lifestyle predispositions, living near transit can *de-generate* vehicle trips. And with the option of carsharing, it might reduce parking demands as well. This has been the case in cities like Zurich, Switzerland. Zurich, for example, has the second highest per capita transit use in the world (over 600 transit trips per capita per year) and the highest per capita carsharing participation anywhere (7% of households are members of Mobility Carsharing Switzerland) [6]. And in spite of having one of the world's highest per capita incomes (on a purchasing power parity basis), only one out of three households has an off-street parking space. Zurich also has the highest commercial real estate prices in Europe (along Bahnhofstrasse) and according to the management consulting firm, Arthur D. Little, ranks number one in quality of life among global cities [6]. In Zurich,

world-class transit, carsharing, parking limits, and prosperity go hand-in-hand.

Another policy response might be the introduction of more flexibility in parking policies for housing near rail stops. Flexibility can be in the form of enabling projects to provide below-code parking levels when justified, e.g., compact projects with short, direct walking connections to transit and perhaps on-site retail establishments. Flexible parking standards provide latitude in providing the optimal number of parking spaces [19]. Flexibility can also take the form of unbundling the cost of providing parking from the cost of building (or renting) housing [20]. This would allow developers to better scale the amount of parking provided to what each tenant or homeowner is willing to pay for each car owned, i.e., let the market demand, rather than a possibly out-dated government fiat, determine supply. And flexibility can be in the form of allowing TOD tenants to choose deeply discounted transit passes for frequent riders instead of a 300 square foot parking space.

TOD Implementation Tools

Moving from the theory to the practice of TOD is often fraught with difficulties, especially in car-dependent countries like the United States. A national survey of urban planners in US cities with TODs cast light on the kinds of implementation tools being used to leverage TOD [4]. Most prominent has been zoning, usually in the form of overlays. Overlay zones are frequently introduced on an interim basis to head-off auto-oriented uses that might compromise a TOD and to specify desired land uses as of right, such as housing and convenience shops. For urban TODs, densities of 20–30 dwelling units per residential acre and floor area ratios (i.e., building area divided by land area) of 1.0 and above are not uncommon. Some of the more progressive TOD zoning districts, such as those found in Portland, Oregon, Seattle, San Diego, and Denver, also lower requirements for car parking and sometimes even for bicycles. The city of San Diego, for instance, recommends parking reductions as high as 15% for urban TODs.

Besides zoning, other tools frequently used in the United States to encourage TOD include: funding for station-area planning and ancillary capital

improvements; density bonuses, sometimes used to encourage affordable housing; and relaxation of parking standards. Next in the order of frequency of usage have been land-based tools, like land purchases on the open market (for land-banking and potential “deal-making”) and assistance with land assemblage. For the most part, redevelopment agencies have applied these tools, meaning their role in leveraging TOD has been mainly limited to economically depressed or blighted neighborhood settings. Because of the higher risk involved, redevelopment tools have often been accompanied by other funding sources, sometimes with a dozen or more participants involved in the process.

Central-City Redevelopment Around Streetcar Lines

A good example of successfully using zoning tools and financial incentives to leverage central-city redevelopment in a transit corridor is the Pearl District in downtown Portland, Oregon. The Pearl District is the most dramatic transformation of downtown Portland in the last 20 years. Once home to a large artist community and an “incubator” for start-up businesses in abandoned warehouses, the Pearl District is now an emerging mixed-use neighborhood of upscale loft housing, parks, art galleries, boutiques, cafes, and restaurants. Since 2001, more than 1,700 condominiums and apartments have been built. Moreover, 55% of all development in downtown Portland has been within a block of the streetcar service since its opening [21].

A major catalyst to the transformation of the Pearl District was the construction of the Portland Streetcar, the first modern streetcar system to be built in the United States. The streetcar has been equal parts housing and transportation tool, as streetcar construction was explicitly linked to high-density development via density bonuses and an innovative developer agreement. As a result of this agreement, the average density of the District is now 120 housing units per acre, the highest in the city. Moreover, properties near the streetcar line more closely approach permissible densities than properties situated farther away [21].

Tax incentives were introduced to moderate possible gentrification effects of streetcar-induced redevelopment. Many affordable housing projects in Portland

receive 10-year property tax abatements if they are within walking distance of a rail line. While the abatements are loosely related to projected price levels and affordability, their primary purpose is to ensure denser development (as specified in the developer agreement) than the market would otherwise support.

Residents in the Pearl District fit the demographic profile found in other Portland area TODs – childless, and either young people seeking smaller lofts, older professionals looking for an urban lifestyle with little upkeep (“downsizing boomers”), or retiring seniors. This variety of homeowner types has contributed to the depth of the market.

The Challenges of Mixed-Use TODs

As dense, mixed-use forms of development, many of the barriers to TOD are generic to all forms of compact growth – not-in-my-backyard (NIMBY) resistance, higher risks and costs, and institutional inertia. Still, some of the barriers to smart growth are more pronounced when it comes to TOD. Mixed-use development is one of them. Mixed land uses, a signature feature of TODs, pose a host of difficulties not only in terms of design but also in lining up funding, investors, and contractors. Planners sometimes impose a design template of ground-floor retail and upper-level housing or offices, i.e., vertical mixing – on any and all development proposals within a TOD. Mixed-use projects are much trickier to design, finance, and sometimes lease than single-use ones [12]. Finding the right formula for mixed land uses can be every bit as difficult as rationalizing parking policies. Vertical mixing is particularly problematic. Quite often, the ground-level retail component of mixed-use TODs suffer the most, in part because they are poorly laid out. Ground-floor retail, for example, is doomed to fail unless it opens onto a street with busy foot traffic and convenient car access. Mixed housing-retail projects also pose unique design challenges. Ground-floor retail needs greater floor-to-floor height (typically 15–18 ft) to be marketable, compared with the 8–10 ft between residential floors. This means the entire ground floor, including multifamily areas, must have higher ceilings, which increases project costs. Ground-floor restaurants pose problems such as where to put the exhaust shafts for kitchens. The exact size and location of restaurant

space may not be known until leases are signed. Designers must thus allow exhaust shafts to be put in several potential locations, which can reduce net leasable space. And ground-floor restaurants might be unappealing to upper-level residences seeking quiet and privacy in the evening. Local governments need to be sensitive to these issues and focus more on achieving a desired land-use mix within a transit station area as opposed to individual parcels, i.e., pursue “horizontal” neighborhood-scale mixes versus “vertical” within-building mixes.

Financing TODs

For TODs to take form, there must be money to pay not only for transit infrastructure, like stations and parking, but also the armature – civic squares, streetscape enhancements, pocket parks, expanded trunk-line sewage and piped water capacity – that surrounds a TOD station. Joint development is the form of value capture that has been most widely used to finance America’s most successful TODs. Under this approach, a public transit agency partners with a private developer to build a real-estate project on land or air rights owned by the transit agency itself. In return, the transit agency receives either revenue (i.e., revenue sharing) or passes on part or all of the costs of rail-station and ancillary construction (i.e., cost sharing). Among the most common forms of revenue-sharing schemes are land leases, air-rights development, and station interface or connection-fee programs. Cost-sharing schemes include sharing construction expenses, incentive-based programs that produce benefits for private financing (e.g., density bonuses), and joint use of equipment like ventilation systems.

As a value capture tool for transit, joint development has the most appeal in settings where a significant amount of land is available to a transit agency, preferably purchased on the open market at a fairly low cost (which usually means prior to formal announcement of a new railway project). This has been the case in Hong Kong where the city’s transit operator, MTR Corporation, obtains more than half of its revenue from ancillary real-estate property development [22]. To the degree joint development is limited to a geographically restricted area, such as one or two small parcels owned by a transit agency, it fails to

capture value from a broader area benefiting from new transit services.

Another common way to leverage and fund TODs, especially in the United States, has been Tax Increment Financing, or TIF. Under this scheme, incremental increases in property tax revenues are channeled back into a district to help revitalize distressed neighborhoods. TIF has been used extensively to encourage TOD in many parts of the United States. In the city of Chicago, half of the 129 TIF districts in 2003 contained a railway station [4]. The state of Pennsylvania applies TIF in Transit Revitalization Investment Districts (TRIDs) to promote economic development and TOD. A criticism of TIF is it diverts tax dollars that would otherwise accrue to a city’s general treasury and as thus creates a privileged (i.e., subsidized) zone. Such equity concerns have limited the application of TIFs to much of the developing world.

Some TODs have relied on impact fees as their chief source of funding. These are charges assessed on new development to defray the cost of expanding and extended public services. In 1981, the city of San Francisco, California introduced a Transit Impact Development Fee (TIDF) levied against new downtown office buildings to produce income for operating and maintaining the city’s MUNI transit network, comprising light rail, diesel bus, trolley bus, and cable-car services. TIDF revenues produce around \$10 million annually. Broward County, Florida, has a Transit Oriented Concurrency system that similarly generates income for transit operations by levying a fixed fee on new development. This program covers around 30% of annual bus-transit operating and capital costs for the county [23]. Because limiting impact fees to new development only can deter investments in all but the healthiest real-estate markets and inflates the cost of housing, it has been applied on a limited basis outside of the United States. In the developing world, exactions are more common wherein developers cover the cost of expanding infrastructure (e.g., water and sewer trunkline capacities) to serve a site.

Future Directions

TODs are a type “smart growth” and sustainable urbanism that have tremendous growth potential due to both supportive public policies and market forces.

Global experiences show that the integration of public transport and land use can yield appreciable sustainability benefits that redound to both the private and public sectors. In many settings, developers know they can earn handsome profits building, leasing, and selling TODs. As long as traffic congestion continues to worsen in TODs and demographic trends continue to shift in the direction of smaller, non-traditional households, there will always be a market demand for TODs. However, forward-looking and pro-active public-sector strategic planning will always be a necessity as well if sustainable and cost-effective TODs are to take form. In many automobile-reliant cities of the world, TODs often take the form of a few nodes, or islands, in a sea of auto-oriented development. Unless there are enough TODs, linearly aligned along natural travel corridors to create a critical mass, few transit ridership (and correspondingly auto-reducing) benefits will accrue. That is, there need to be enough origin-destination combinations formed by TODs in non-downtown settings to draw enough motorists into the trains and buses so as to put an appreciable dent in car traffic. Otherwise, stand-alone TODs can very much backfire by creating dense nodes in otherwise auto-oriented settings so as to increase traffic congestion along roads feeding into transit stations.

What is likely on the horizon for particularly progressive-minded cities might be the emergence of “Green TODs.” Green TOD is a marriage of TOD and Green Urbanism. This combination is, potentially, a potent elixir. TOD works on the VMT-reduction side of shrinking a city’s carbon footprint. Green Urbanism reduces emissions and waste from stationary sources, in the form of green architecture and sustainable community designs. In combination they can deliver energy self-sufficiency and zero-waste living. Renewable energy might come from solar and wind as well bio-fuels created from human waste and other byproducts. Recycling and reuse of materials, community gardens, and low-impact building materials further reduce the carbon footprint of Green TODs (Table 1).

Synergies would like accrue from combining TOD and Green Urbanism. The higher community densities needed to fill the trains and buses that serve TODs at the same time reduce heating and cooling expenses from the embedded energy savings of shared-wall construction. Smart electric grids generated by photovoltaic

Transit-Oriented Development and Land Use. Table 1
Forms of emission reductions form green TODs

TOD	Green urbanism
Mobile sources • <i>Transit design:</i> World-class transit (trunk and distribution) Station as hub • <i>Non-motorized access:</i> (bike paths, pedways) • <i>Bikesharing/carsharing</i> • <i>Minimal parking:</i> (reduced land consumption, building massing, and impervious surfaces) • <i>Compact, mixed uses</i>	Stationary sources • <i>Energy self-sufficient:</i> (renewably powered – solar, wind turbines) • <i>Zero-waste:</i> (recycle, reuse, methane digesters, rainwater collection for irrigation, and gray water use) • <i>Community gardens:</i> (compost, canopies) • <i>Buildings:</i> Green roofs Orientation (optimal temperatures) Materials (recycled, low impact)

panels and wind turbines designed as canopies above stations might also power light-rail cars, plug-in hybrids, and electric buses. Additionally, the compact, mixed-use pattern of development around rail stations can support walking, biking, and the use of limited range private vehicles to reach relatively nearby destinations, thus helping to grow such infant-industries as electric vehicles. One could imagine a future of hydrogen-fueling and electric-battery swap depots in a green community focused on a rail node and for which the destinations of many travelers are close by.

Hammarby Sjöstad, a brownfield redevelopment in the city of Stockholm, is an example, *par excellence*, of marrying of TOD and green urbanism. The combination of railway services, carsharing, and bike-sharing has dramatically reduced vehicle-kilometers traveled of Hammarby’s residents and correspondingly greenhouse gas emissions and energy consumption. In 2008, 25% of trips by Hammarby’s residents were by bicycle or walking and 52% were by transit, much higher than regional averages [24]. Carbon dioxide emission from private cars was 52% less for households in Hammarby Sjöstad than in planned communities of similar income structure in the suburbs of Stockholm [24]. And the design of an energy self-sufficient and low-waste community has further shrunk the project’s environmental footprint.

Today, residents of Hammerby Sjöstad produce 50% of the power they need by turning recycled wastewater and domestic waste into heating, cooling, and electricity.

The future of TOD, worldwide, is undeniably bright. As long as urban ills like traffic congestion and air pollution – and global concerns like oil dependency and climate change – persist or worsen, TOD will be embraced as a desirable form of urbanism. Rapid urbanization, combined with equally rapid urban and inter-city rapid rail construction, in countries like China could mean more TODs find particular acceptance in former third-world countries that are rapidly industrializing and modernizing.

Bibliography

1. Cervero R (2008) Transit-oriented development in America: strategies, issues, policy directions. In: Haas T (ed) *New urbanism and beyond: designing cities for the future*. Rizzoli, New York, pp 124–129
2. Renne J (2009) From transit-adjacent to transit-oriented development. *Local Environ* 14(1):1–15
3. Urban Land Institute, San Francisco (2008) Bay area TOD marketplace. Urban Land Institute, San Francisco, p 1
4. Cervero R et al (2004) Transit oriented development in America: experiences, challenges, and prospects. TCRP report 102. National Academy Press, Washington
5. Cervero R, Ferrell C, Murpy S (2002) Transit-oriented development and joint development in the United States: a literature review, research results digest number 52. Transportation Research Board, Washington
6. Cervero R (1998) *The transit metropolis: a global inquiry*. Island Press, Washington
7. Kenworthy J, Laube L (1999) *An international sourcebook of automobile dependence in cities: 1960–1990*. University of Colorado Press, Boulder
8. Consejo Nacional de Política Económica y Social (2000) *Sistema de Servicio Público Urbano de Transporte Masivo de Pasajeros de Bogotá*. Departamento Nacional de Planeación, Documento No. 3093
9. Targa F, Rodríguez D (2004) Analysis of Bogotá's bus rapid transit system and its impact on land development. *Carolina Plann J Winter*:26–36
10. Rodríguez D, Mojica C (2009) Capitalization of BRT network expansions effects into prices of non-expansion areas. *Transp Research Part A* 43(5):561–575
11. Cervero R, Sarmiento O, Jacoby E, Gomez L, Neiman A (2009) Influences of built environments on walking and cycling: lessons from Bogotá. *Int J Sustain Transp* 3:203–226
12. Dunphy R, Cervero R, Dock F, McAvey M, Porter D, Swenson C (2004) *Developing around transit: strategies and solutions that work*. Urban Land Institute, Washington
13. Cervero R (1994) Transit-based housing in California: evidence on ridership impacts. *Transp Policy* 1(3):174–183
14. Lund H, Willson R, Cervero R (2006) A re-evaluation of travel behavior in California TODs. *J Arch Plann Res* 23(3):247–263
15. Cervero R, Arrington G (2008) Vehicle trip reduction impacts of transit-oriented housing. *J Public Transp* 11(3):1–17
16. Cervero R (2007) Transit oriented development's ridership bonus: a product of self selection and public policies. *Environ Plann A* 39:2068–2085
17. Bernick M, Cervero R (1997) *Transit villages for the 21st century*. McGraw-Hill, New York
18. Daisa J (2004) Traffic, parking, and transit oriented development. In: Dittmar H, Ohland G (eds) *The new transit town: best practices in transit-oriented development*. Island Press, Washington, pp 114–129
19. Cervero R, Golub A, Nee B (2007) City carShare: longer-term travel-demand and car ownership impacts. *Transp Res Rec* 199:70–80
20. Dunphy R, Myerson D, Pawlukiewicz M (2004) Chapter 7: Ten principles for developing around transit. In: Dunphy R et al (eds) *Developing around transit: strategies and solutions that work*. Urban Land Institute, Washington
21. Shoup D (2005) *The high cost of free parking*. Planner's Press, Chicago
22. Cervero R, Murakami J (2009) Rail + property development in Hong Kong: experiences and extensions. *Urban Stud* 46(10):2019–2043
23. Fogarty N, Eaton N, Belzer D, Ohland G (2008) *Capturing the value of transit*. Center for Transit Oriented Development, Oakland
24. Brick K (2008) Summary report: follow-up of environment impact in Hammarby Sjöstad. Grontmij AB, Stockholm

Transmission Blackouts: Risk, Causes, and Mitigation

DAMIR NOVOSEL

Quanta Technology, LLC, Raleigh, NC, USA

Article Outline

Glossary

Definition of the Subject and Its Importance

Introduction

Grid Development History

Challenges in Taming the Grid from Wide-Area Blackouts

History and Examples

Pre-outage Conditions and Risks for Blackouts

Symptoms of Blackouts
 Power System Modeling and Analysis
 How Disturbances Turn into Blackouts
 Blackout Prevention
 System Integrity Protection Schemes
 System Restoration
 Future Directions
 Conclusions
 Acknowledgment
 Bibliography

Glossary

DOE Department of Energy
EMS Energy Management System
FACTS Flexible AC Transmission System
FERC Federal Energy Regulatory Commission
GPS Global Positioning Satellite
ISO Independent System Operator
NERC North American Reliability Council
PDC Phasor Data Concentrator
PMU Phasor Measurement Unit
RAS Remedial Action Scheme
TSO Transmission System Operator
SCADA Supervisory Control and Data Acquisition
SIPS System Integrity Protection Scheme
UCTE Union for the Coordination of Transmission of Electricity
WAMPAC Wide Area Monitoring, Protection, and Control
WECC Western Electricity Coordinating Council

Definition of the Subject and Its Importance

Power system blackouts result in complete interruption of electricity supply to all consumers in a large area. While it may be possible to trace a blackout's beginning to a single incident (e.g., transmission line sagging into a tree), cascading outages are the result of multiple low-probability events occurring in unanticipated or unintended sequence. The likelihood of power system disturbances escalating into a large-scale cascading outage increases when the grid is already under stress.

Blackouts may look like bad luck, but they are result of how the grid is managed. Statistically, a sequence of low-probability contingencies with complex interactions causing a blackout may not happen often but

will eventually take place unless measures are taken to prevent it. Presently, some of the grids around the world (e.g., parts of the North American grid) may be susceptible to blackouts as aging and insufficient grid infrastructure may not be adequate to accommodate grid changes, such as renewable generation resources and load growth. Deployment of "Smart Grid" monitoring, control, and protection devices, software tools, and telecommunication infrastructure helps better manage the grid, but does not replace investments in the infrastructure.

Although large-scale blackouts are very low-probability events, they carry immense costs for customers and society in general as well as for power companies. It is easy to misjudge the risks and costs of such extreme cases.

Although widespread outages cannot be completely prevented, their occurrence can be reduced, propagation (size) and consequences arrested, and restoration sped up. In last couple of decades, power systems around the world have been more stressed as the capacity reserves and system margins have been reduced resulting in more blackouts with huge costs to the society.

Introduction

Wide-area electrical blackouts have raised many questions about the specifics of such events and the vulnerability of interconnected power systems [1–4]. Power systems are complex interconnected machines with many components operating in harmony when the system is balanced. When a problem occurs in part of the system, the impact of the trouble may cause to the system losing its balance momentarily. In most cases, the system is immediately isolated and shortly after recovers with no further propagation observed outside of the immediate area.

Exchange of information stemming from the recent worldwide blackout findings and "smart grid" technology innovations shed new lights on the current conditions and needs of power systems. Examination of the root causes, the resulting effects on neighboring systems, and implementation of proven solutions to help prevent propagation of such large-scale events should help in designing and operating reliable "smart grid" and power delivery infrastructures for today and in the

future. Power system professionals can take in consideration the costly lessons of the past, maintain a library of historical lessons about “What and Why it Happened” for the generations to come, and act as catalysts to achieve desired level of power system reliability.

The high cost and the need for extensive mitigation strategies against grid congestions, combined with this type of probabilistic assessments, have led into risk management not focusing on appropriate, cost-effective mitigation actions. From a broader perspective, a misconception may be formed about the grid reliability or its exposure to large-scale outages every so often. It is easy to misjudge the risk of such extreme cases. (Risk is the product of the cost and associated probability, and both factors are very hard to assess accurately.) The high costs of extensive mitigation strategies (e.g., building new transmission lines), combined with inaccurate probabilistic assessments (“blackouts will not happen in my system”), have led to inadequate risk management practices, including not focusing on cost-effective prevention and mitigation initiatives. Such initiatives can provide value through avoidance of huge blackout costs. The following stakeholders benefit from outage/blackout avoidance:

- The society/ratepayers. For example, society costs for August 14, 2003 blackout in the USA and Canada and for August 2006 WECC blackout were estimated at \$7B and \$1B, respectively.
- The electric utilities, ISOs, generation producers, etc. Costs of service restoration, undelivered energy, cost of litigation, and the negative impact on stock price.

Understanding the complexities of the interconnected power grid, need for proper planning, good maintenance, and sound operating practices are the key to deliver electric power to modern day necessities and prevent the problems of tomorrow’s grids. This article offers practical insight on risks and leading causes of widespread blackouts and how best to prevent them to achieve higher levels of grid reliability.

Grid Development History

In the 1930s, electrical power was delivered to consumers around the world by individual, not connected

electrical systems. Regional systems have been created next to make power systems more robust and delivery more reliable. The original function of the interconnected systems was to form the backbone for the security of supply and to reach its required high reliability level at reasonable costs. In the 1950s, the power system professionals foresaw the importance of further improving delivery of reliable electricity to consumers. The grid systems have been developed to assure mutual assistance between transmission system participants and/or national subsystems including common use of reserve capacities to optimize the use of energy resources by allowing exchanges between the systems.

Thus, a strategy of interconnecting neighboring systems to improve reliability and security margins became a reality around the world. Coordinated rules for the mutual support of interconnected systems were defined and adopted by the power pool members between power systems, including interconnections among countries (e.g., UCTE in Europe). Since the late 1970s, the electrical transnational infrastructures were exploited more and more for energy exchanges that took advantage of the different production costs of electricity in the various nations or interconnected grids in order to deliver lower cost energy and achieve maximum profits. However, the bulk power system was not originally engineered to transfer large amounts of power between neighboring systems over long ranges but to enable neighboring utilities to support each other during stressed conditions.

In recent years, deregulated energy concepts have created additional burden on the system for using the same conductor for additional capacity increasing already high transfers as new generation sources (including renewable energy) are brought on-line to meet the demand. The high level of power exchanges in today’s energy market is technically being provided outside of the scope of the original system design. The higher demand, coupled by low level investments in technology and infrastructure upgrades and capacity increase, has led the Control Area operators to run the system close to the edge, as close to the limits as permitted by the reliability criteria and, sometimes, beyond the limits. At the same time, our increased respect and awareness for the environment and the “Not in My Backyard” sentiments have made it difficult

to site transmission lines or major local generation sources, especially in the more densely populated areas, where load is heavy. These difficulties make the system expansions expensive and difficult, and offer new challenges to deliver reliable power.

Recently, major investments have been made worldwide in “smart grid” technologies, e.g., US DOE investment grants. Major motivation for Transmission “smart grid” is in preventing blackouts using Wide Area Monitoring Protection and Control (WAMPAC) technologies, such as synchronized measurements using GPS signals.

There are still isolated power systems around the world (e.g., island networks). Unlike interconnected grids, those power systems do not have flexibility to get help from the neighboring systems to optimally balance generation and load, both during normal and stressed conditions. Those systems are more vulnerable to generation and transmission system outage and require long-term, coordinated planning and investments in the complete electrical grid infrastructure to achieve reliable and cost-effective grid operation and maintenance.

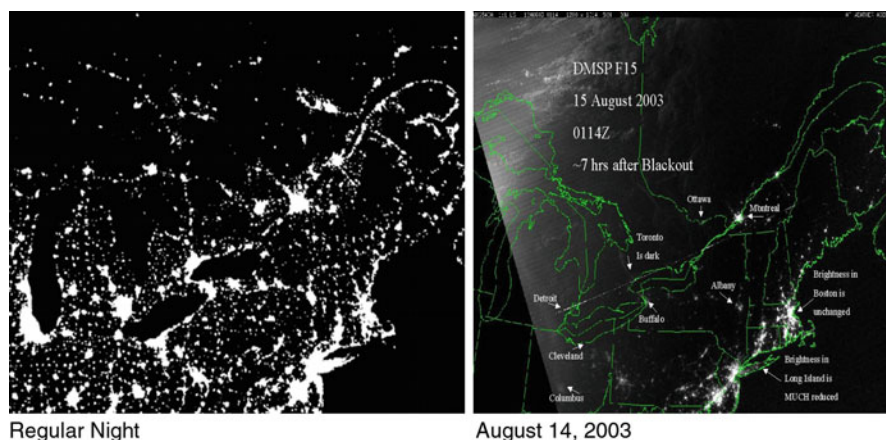
Challenges in Taming the Grid from Wide-Area Blackouts

Evolutions in technology continue to improve all aspects of our lives from precision surgical equipment

for use in highly critical operations, to automatic banking anytime of the day, electric rail systems to reduce emissions, and to our home appliances. Many of the technological innovations have been achieved in recent quarter century and have readily found their way in our daily lives. The modern day amenities and our respect for the environment have also increased our dependency on energy, hence, our expectations for uninterrupted reliable power. However, the demand for availability of power for many of these modern day equipment has not been systematically and uniformly considered. Modern technology is the catalyst driving power delivery demanding grid reliability, and the marked increased dependency on availability has raised the bar on human expectation.

Some of the electrical power systems that experienced blackout in last decade (e.g., North America, Europe) are among the most reliable systems worldwide. However, they are subject to a host of challenges – aging infrastructures, renewable and distributed energy integration, transmission and distribution grid expansion to meet the growing load demand and transfer renewable energy to load centers, etc. [Figure 1](#) shows impact of the 2003 blackout on supplying power to the consumers (before and after the blackout).

After some major blackouts worldwide in the last 2 decades, utilities have hardened their electrical systems, and regulatory organizations (e.g., NERC and FERC and state regulators in the USA) have focused on



Regular Night

August 14, 2003

Transmission Blackouts: Risk, Causes, and Mitigation. Figure 1
Effects of August 14, 2003 blackout in North America

defining and enforcing reliability standards. This has resulted in better planned, operated, and maintained grid. However, one of the challenges facing the electric industry today is the balance between reliability, economics, environmental, and other public purpose objectives to optimize transmission and distribution resources to meet the demand. Resources and transmission adequacy are necessary components of a reliable and economic supply. Though the reliability and market economics are driven by different policies and incentives, they cannot be separated when the objective is reliability and availability. Today, grid planning in the regional and interregional environment faces an extremely difficult task given the challenge to achieve resource adequacy in today's restructured industry, where market economics, local concerns, and renewable energy resources often derive the decision for generation facility siting remote from major load centers. Equally difficult is planning for an adequate transmission system when the location of future generation facilities is uncertain and the lead time for transmission construction is very long (just the permitting process may take several years).

It is more important than ever to find ways to project transmission and distribution growth, solutions to deploy, and criteria to be applied to guide prudent investment decision. Some of the key areas to address are the following:

- Integration of renewable energy – As renewable resources (e.g., wind and solar) are located far from the load centers, additional stress was put on the system, causing higher vulnerability to outages.
- Price for reliability – Costs and risks transmission owners and customers are willing to assume. The power industry is accustomed to optimize investments and evaluate return on investments based primarily on financial aspects of trading energy and serving load within certain reliability criteria. This is often done without considering financial aspects of unavailable energy (undue service interruptions) due to low reliability and slow restoration that incurs significant costs to the society. This is an incomplete financial model that results in suboptimal investment strategies.
- Large regional geographic areas should be included in the scope of transmission planning and

decision-making. However, when assets need to be built, it is not easy to identify true beneficiaries and how costs are to be shared.

- Quick restoration – As it may not be possible to completely avoid outages, it is required with today's technology to reduce power restoration.

Electricity is the key resource for our society; however, strategic planning (regional and national) requires additional priority to improve system reliability.

History and Examples

Within the last decade, the number of wide-area outages has rapidly increased (blackouts in north-east USA and Canada, western USA, Brazil, Italy, Sweden, Denmark, England, India, Malaysia, Australia, New Zealand, Greece, etc.) affecting over 300 million customers worldwide. As the likelihood of low-probability events escalating into a cascading outage increases when the grid is already under stress due to preexisting conditions, one can conclude that power grids are more prone to disturbances than ever.

History has showed that both unscheduled and scheduled (regular maintenance) outages have affected power system's balanced operation hence signifying the grid complexity during managed conditions [1, 2, 5–7]. In the case of the August 1996 North America disturbance that affected 12 million people [5], a series of equipment were being removed for maintenance in parts of the Western grid when the weather was moderate, yet these pieces of equipment were needed to support transfers in other parts of the Western grid, which was experiencing extremely high temperatures.

Analysis of the 2003 incident in the north-eastern part of the USA [6], which affected 50 million people, revealed that a series of cascading events over the course of several hours, rather than a single instantaneous problem, initiated the major disturbance phenomenon that toppled the large part of the US Eastern Interconnection. The blackout itself was preceded by scheduled generation outages and line tripping caused by overgrown trees in the right-of-way. It was accompanied by failures of EMS/SCADA alarm systems, which prevented operators from diagnosing problems. The initial line disconnection caused another 345 kV line to overload and sag into a tree, resulting in

remaining 345 kV lines to disconnect. Underlying 135 kV lines overloaded and disconnected, followed by tripping of generating units, causing a part of the system to go black. A lack of communication and coordination between utilities and ISOs involved exacerbated the problem. The grid was restored in 1–2 days (depending on the area affected and ability to get generators back on line).

The September 2003 disturbance in the connected European grid affected 57 million people. Several 220–400 kV lines were out of service for maintenance reasons prior to the event occurrence. Italian grid was importing 6 GW from the rest of the grid. One of the 380 kV lines disconnected due a tree contact. The parallel line overloaded, and although the import was reduced, it was not enough to prevent sagging into a tree and disconnecting. Other lines to Italy overloaded and tripped, resulting in isolating Italian grid 12 s after the loss of the second line. During those 12 s, low voltage in northern Italy caused generators to disconnect. Other countries tripped generation (app. 6.7 GW). In summary, 2.5 min after islanding, Italy goes black separated from the rest of the grid. However, unlike the blackout in the USA, it took only 5–9 h to restore the power to major cities. Main reasons are for faster restoration was type of generation that was able to get back on line faster and accompanying restoration processes (e.g., so-called black-start capabilities allowing fast generation reconnection).

The most severe disturbance in the history of UCTE considering number of Transmission System Operators (TSOs) involved happened in October 2006 [7] when system separated in three islands. Following were key reasons for the disturbance:

- N-1 criterion (planning criteria so that a single equipment outage does not create a system disturbance) was not fulfilled as wind generation was not adequately predicted. Large wind storms increased production in northern Germany.
- Inappropriate inter-TSO coordination and training. The time of the planned line outage was changed, and new conditions were not checked.
- Uncoordinated protection settings on both sides of the critical line.
- Uncoordinated operation of wind generators.
- Inadequate coordination of restoration.

All the above blackouts included combination of phenomena such as line overloads, voltage and angular instability, and system separation. Those blackouts have served as catalysts in propelling the power industry toward analyzing blackouts and finding solutions to prevent such occurrences. The 1994–1996 blackouts in the western part of North America resulted in formation of investigating teams which ultimately made a combined list of 130 conclusions and 54 recommendations [7]. The August 14, 2003 investigating teams identified 60 recommendations (14 by North American Reliability Council, or NERC, and 46 by the US Canadian Task Force) [6]. The Union for the Coordination of Transmission of Electricity (UCTE) identified 14 observations for the September 28, 2003 outage, very similar to those identified for the outages in North America that happened just a month before. Although major improvements have been accomplished after the blackouts in corresponding grids, the similarities in findings among those blackouts prove that if recommendations had been used among grids, the impacts of the outages could have been minimized [4].

By some accounts [8]:

- These kinds of outages are consistent with historical statistics, and they will continue to happen. Some may compare blackouts to events such as long-term weather forecasting (hurricane and mudslides) or to natural disasters such as earthquakes as being difficult to predict and/or to prevent.
- The more immediate problem may be the industry's investment of less than 0.5% in R&D, one of the lowest rates for any industrial sector.
- A power system is composed of hundreds of thousands of pieces of equipment from bulk autotransformers, high-voltage transmission systems, to a light bulb. It has been suggested that one could not get a computer big enough to model a complex system as, for example, the Eastern Interconnection and perform the planning studies.
- Large blackouts occur because the grid isn't forcefully engineered to prevent them. Purposely weakening the grid can reduce large blackouts but would increase the frequency of smaller ones.

Chifong Thomas, formerly transmission planning engineer from Pacific Gas and Electric Co. points out that if large blackouts occur anyways, as they are hard

to predict as earthquakes, then nothing anyone could do would make any difference. This contradicts the premise that industry underinvestment in R&D is the immediate problem. If the first premise is true, the second one is irrelevant. She also points out that if the grid planning is done comprehensively and correctly, then operators with proper training should have time to respond to contingencies and/or to limit the impacts of the contingency.

Dr. Mayer Sasson, principal advisor of Electric Markets Policy Group at Consolidated Edison Co. of New York, highlights the vulnerability of the interconnected grid by pointing out that it may take only one member of a control area to be operating outside of the reliability limits for any reason, including unintentionally, to cause a cascading outage that can quickly propagate to neighboring interconnected system control areas. This underscores the criticality of complying with mandatory reliability rules.

Dr. Bogdan Kasztenny, power system protection and control application manager at Schweitzer Engineering Laboratories, raises the important issue of human factor involvement in the last few critical hours before any major event. “Even in a poorly designed and under-invested system, operators could sometimes salvage an event that seems to be disastrous, or collapse an otherwise quite secure situation in a strong and reliable system. It is not so with earthquakes. Power systems are man-made creations run by humans. Operational procedures, availability and accuracy of real-time information, adequate training including dry runs on simulators are technical means that improve response of the operators.”

Kasztenny emphasizes that the power system is a complex generating, transmitting, and distributing system for a medium that cannot be stored, or buffered, in the reality of very limited redundancy. This makes it quite different from the banking, phone, or similar systems. Other systems are subject to brownouts. However, the power system must balance the medium under physical constraints and is therefore subject to collapse if the constraints are violated.

Pre-outage Conditions and Risks for Blackouts

The grid is a tremendously complex system, and the interconnections that allow us to benefit from higher

reliability and lower costs also cause the domino failures experienced in many parts of the world in recent years. Although there is a tendency to point at one or two significant events as the main reasons for triggering cascading outages, major blackouts are typically caused by a sequence of low-probability multiple contingencies with complex interactions. The three “Ts” – Trees, Tools, and Training – have been identified as the leading focus areas to prevent widespread outages not caused by natural events. However, disturbances have occurred following extremely low-probability successive unscheduled equipment outages beyond planning criteria. There have also been cases of system disturbances caused by scheduled equipment outages when the electrical system has not been adjusted, for continued safe operation, prior to the equipment being removed. Low-probability sequential outages are also not anticipated by system operators, thus rendering the power system more susceptible to wide-area blackouts. As the chain of events at various locations in the interconnected grid unfolds, operators cannot act quickly enough to mitigate the fast developing disturbances.

Power systems are engineered to allow for reliable power delivery in the absence of one, two, or more major pieces of equipment such as lines, transformers, or bulk generation, commonly referred to as contingency conditions. For example, after the 2003 blackout in North America, North American Electric Reliability Council (NERC) set forth the reliability standards and performance requirements that are enforced by audits. However, the complexity of the grid operation makes it difficult to study the permutation of contingency conditions that would lead to perfect reliability at reasonable cost. Accurate sequence of events is difficult to predict, as there is practically an infinite number of operating contingencies. With system changes, e.g., independent power producers selling power to remote regions, load growth, new equipment installations that cause significant changes in power flow to name a few, these contingencies may significantly differ from the expectations of the original system planners and engineers.

The likelihood of power system disturbances escalating into a large-scale blackouts increases when the grid is already under stress due to preconditions. Those preconditions are summarized as follows:

- Congested grid with tight operating margins
 - Not building lines or generation as fast as required, exacerbated by difficulty in identifying business models to recover costs, and a cumbersome permitting process
 - Not well-planned wholesale merchant transactions with scheduled transactions not changed to allow for transmission relief when required
- System stress caused by intermittent renewable energy and/or suboptimal operating practices
- Insufficient reactive support where and when required to maintain required voltage levels
 - Adequate dynamic reactive power is required close to the load as reactive power cannot be transferred over long distances
- Uncoordinated planning between transmission and generation
 - Inadequate system reserve, such as generation spinning reserve
- Inadequate planning/operation studies
 - No routine use of an effective contingency analysis tool
 - Uncoordinated interregional transmission planning
- Aging infrastructure, prone to failures, accompanied by insufficient level of investment in maintaining the grid
 - It is more and more difficult to isolate and remove equipment for maintenance
- Both scheduled and uncoordinated maintenance
- Lack of system and component knowledge (e.g., system operator not aware of line loading margins)
- Inadequate right-of-way maintenance or environmental policies versus right-of-way vegetation management
- Weather (high temperatures; wind, thunderstorm, fog, etc.)
- Regulatory uncertainty
- Inadequate Automatic Warning, Protection, and Control Systems

In general, combination of various factors makes power systems more susceptible to disturbances.

Symptoms of Blackouts

It is the cascading events that cause disturbances to propagate and turn into blackouts. System is stressed

and as system and equipment faults occur, the chain of events starts. For example, some generators and/or lines are out for maintenance, line trips due to a fault. Other lines get overloaded, and another line gets in contact with a tree and trips. There is a hidden failure in the protection system (e.g., outdated settings or HW failures) that causes another line or generator to trip. At that stage, power system is faced with overloaded equipment, voltage instability, transient instability, and/or small signal instability. If fast actions (e.g., load shedding, system separation) are not taken, system cascades into a blackout.

Evaluation of disturbances shows that protection systems have been involved in 70% of the blackout events [9]. For example, distance relays trip on overload and/or low-voltage sensitive or ground overcurrent relays trip on high unbalance during high load. Inadequate or faulty alarm and monitoring equipment, communications, and real-time information processing can further exacerbate disturbances in the system. Either information is not available or operators are flooded with alarms, so they cannot make proper decisions fast.

Human error and slow operator response are major contributing factors for cascading outages. As a disturbance develops, operators in various regions are faced with the questions “is the best course of actions to sacrifice own load, cut interties, or get support from neighbors?,” “should we help or should we separate?.” Important aspect in designing connected power systems is that individual systems should not allow cascading outages to spread throughout the system.

There are a number of other contributing factors that allow a blackout to spread, including lack of coordinated response among control areas. If each region focuses primarily on its own transmission system, the total connected system may not be reliable. As it is very difficult for operators to decide on best actions during a fast developing disturbance, it is desirable to take automated actions before system separates or to separate it in a controllable manner.

Generally, disturbance propagation involves a combination of several phenomena:

- Equipment tripping due to faults or overloads (e.g., transmission lines and transformers). These events

may cause other equipment to get overloaded, creating a cascading event contributing further to system-wide outages.

- Power system islanding (frequency instability) when power system separates. Islands are formed, with an imbalance between generation and load, causing the frequency to deviate from the nominal value, leading to additional equipment tripping.
- Loss of synchronous operation among generators (angular or out-of-step instability) and small signal instability that may cause self-exciting inter-area oscillations if not damped.
- Voltage instability/collapse problems that usually occur when the power transfer is increased and voltage support is inadequate because local resources have been displaced by remote resources without the proper installation of needed transmission lines or voltage support devices in the “right” locations.

Table 1 shows the types of wide-area disturbances likely to occur in two different types of interconnected power grids, namely, meshed network versus an interconnected transmission system of narrow corridors consisting of extensive generation tied to the interconnection [10]. One star indicates that those phenomena

did not affect a particular grid configuration in the past, but have affected modern grids. Two stars indicate major phenomena in a particular grid configuration.

The characteristics of the power system influencing the types of mitigation methods have been described in a variety of literature [1–4, 9–13]. The relative time of action for different type of events, from normal to extreme, varies depending on the type and speed of the disturbance and the need for coordination. Example of the time line for different type of events is shown in Fig. 2.

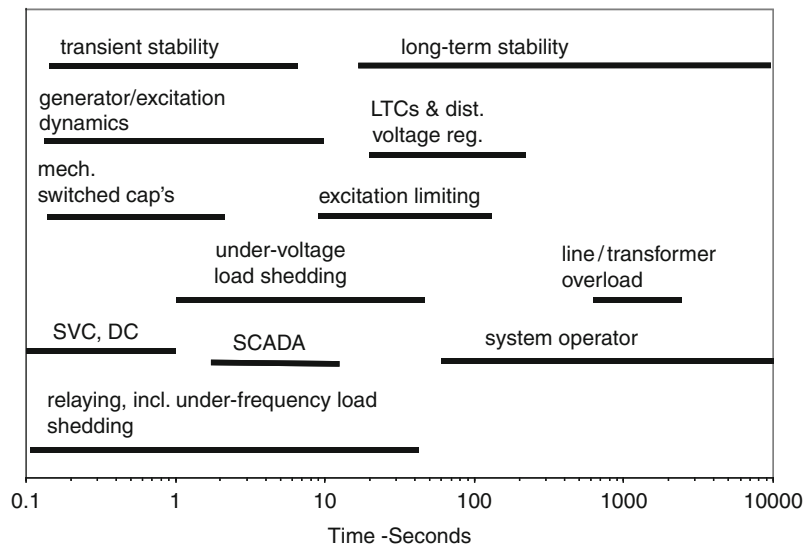
Deployment of a well-coordinated overall defense plan to prevent blackouts requires implementation and coordination of various schemes and actions, spanning different time periods. When events can be controlled to result in gradual shutdown of the available resources rather than permitting them to cascade, their impact could be minimized. A practical approach to identifying stressed system conditions, which are symptomatic of cascading, is feasible. This can be done by defining and modeling the power system parameters considered for reliable operation such as voltage, frequency, or phase angle at critical locations. Such findings can then be implemented through investments in Hardware equipment and Software tools that help manage the grid more effectively.

Transmission Blackouts: Risk, Causes, and Mitigation. Table 1 Types of wide-area events for different transmission systems

System config.	Densely meshed power system with dispersed generation and load		Lightly meshed transmission systems with localized generation and load	
Events	Located in a large interconnection	Not interconnected or by far the largest partner	Located in a large interconnection	Not interconnected or by far the largest partner
Overloads	**	**	*	*
Frequency instability	*	**	*	**
Voltage instability	*	*	**	**
Transient angle instability	*	*	**	**
Small signal stability	*	*	*	*

*Phenomena did not affect a particular grid configuration in the past, but have affected modern grids

**Major phenomena in a particular grid configuration



Transmission Blackouts: Risk, Causes, and Mitigation. Figure 2

Some time frame factors of power system dynamics

Power System Modeling and Analysis

Electrical grids have been characterized as the most widespread interconnected, complex, dynamic systems made by humans. A power system carries tremendous amount of electricity that all depend on. Although blackouts are difficult to predict and prevent, the notion that it is not possible to simulate the grid behavior, to identify and address most vulnerabilities, is not accurate.

Carson Taylor, power system simulation expert (formerly with Bonneville Power Administration), confirms that today's technology allows for detailed modeling of complex power systems. Taylor emphasizes that the Eastern Interconnection has been simulated in great detail with 40,000+ bus models. Simulations with 100,000+ models are feasible. Feasibility increases with IT advancements. Computation is not a large problem. Computation is scalable by parallel computation on multiple servers; cases can be farmed out to the multiple servers. Simultaneous processing method is used for energy management system dynamic security assessment.

Professor Göran Andersson, Eidgenössische Technische Hochschule (ETH) University in Zurich and a noted expert in power system dynamics, concurs with Taylor and brings up an important issue that the

challenge is not technical, i.e., modeling or computer capacity related; rather it is the data management and proper interpretation. Professor Andersson also underlines that the "Statistical methods can give some information and insight concerning the sizes and frequencies of blackouts, which is of value. However, the problem is the calibration of the frequency scale, since this requires a detailed modeling of the interactions. The statistical insights, however, do not provide to any large extent the guidance to avoid blackouts to the power system professionals operating the systems. In terms of modeling, the details of the system and the interactions between different components and sub-systems are indispensable."

Both Andersson and Taylor emphasize that best practice is to continually improve and update models and compare simulations with real power system response. This is not unique for power systems and is continuously being done in the industry and academia. In addition, much progress has been made over the years, and even if predictions from simulations are problematic, efforts to model, simulate, and validate performance provide invaluable insight for a reliable power system operation.

In conclusion, the power grid can be modeled and studied; however, as there is an infinite number of

contingencies that can occur and the current state is not precisely known, it is not possible to exactly predict disturbance propagation far in the future. However, innovations in power system tools from planning, to monitoring, to operation are strategies that would help in meeting the challenges of the twenty-first century expectations in reliable power delivery.

How Disturbances Turn into Blackouts

It has been demonstrated time and again that wide-area blackouts are caused by operation of the interconnected power system outside of the operating limits or for operating conditions that have not been thoroughly studied. As described by Vahid Madani, principal engineer at PG&E, one of the ways to understand the challenges of power delivery is to liken it to driving the highways via motor vehicle. As motor vehicle operators, most of us have experienced a wide variety of unplanned difficulties and challenges slowing down our travels. In many ways, widespread outages have similar characteristics as a highway traffic gridlock. Some of the key aspects of traffic jams include the following:

- Not enough lanes to accommodate the growing demand.
- As one or two lanes are closed, the traffic flow is significantly impacted, and in the most severe case, it causes a complete highway shutdown. A circumstance out of control of those not involved with the accident, yet they are stuck (blackout).
- A properly designed system would include major alternate freeways that are easily accessed through proper detours to minimize impact from the gridlocks.
- Repair of the old highways may not be enough to solve traffic congestion. At times prudent investments in building new traffic lanes, or completely new thruways, and innovative traffic control systems are needed.

Dr. Daniel Karlsson, system analysis and protection expert from Sweden, also compares the expansion of highway traffic with an expansion of the power grid as both have grown from minor systems to extremely important infrastructures in the same time period and both continue to grow. However, he points out

that the public accepts a much higher degree of failures (e.g., people seriously injured or worse) due to automobile traffic than they accept a widespread power outage, demonstrating the important role of electricity in our daily lives and hence the criticality of reliable power delivery.

- Karlsson brings up a historical perspective to explain how the large interconnected power systems of today have been formed due to constant demands on capacity, economy, and reliability. He points out that initial power systems were small grids with low capacity and low reliability. Extensions emerged as dictated by capacity needs. To improve reliability of the individual systems by allowing support from the neighbors, these systems were interconnected. New phenomena appeared, such as transient instability, resulting in studies and actions to counteract them. Complex powerful power systems introduced new transmission capacity problems and actions to counteract them (series capacitors inserted in the transmission line, shunt capacitors and reactors, automatic tap changers to control transformer voltage, etc.). This consequently brought new phenomena like sub-synchronous resonance and voltage instability and, again, a need for new studies and equipment, such as Flexible AC Transmission Systems (FACTS), High Voltage DC (HVDC) links. Dr. Karlsson emphasizes that the complexity of the present power system, and factors such as environmental and rights of way, governmental, cost-to-benefit evaluation for large scale investments will derive us to maximize the use of current assets without truly addressing the rising demand for new infrastructure. This trend will continue to challenge the industry with new technology and actions: superconductivity, energy storage, micro-grid, etc.

There is also a strong relation between system size and reliability. In the ever growing demand to be supported by the power system, our industry tends to push the limits challenging the reliability with a question, “where is the edge?” This is particularly true in many places such as North America, where the expansion efforts are not growing as much, or as rapidly as, the grid is being utilized.

It is noted that weakening or splitting the grid on purpose during normal operation is not a sound alternative, as this strategy would make a full circle for the

grid. For example, by operating the power grid in separate islands, one could easily weaken the power system. That would effectively mean coming back to the original design of the small grid with low capacity. Separating the grid into islands would also impact deregulation, e.g., each DG would be selling power to its own neighborhood. The small utility will soon need occasional support for reserve margin, voltage, and reactive margins from neighboring systems, hence, back to the interconnected grid. The solution is not in separating the grid into islands, rather to resolve transmission problems to mitigate potential for widespread cascading outages.

Chifong Thomas dismisses the argument that large blackouts occur because planning engineers spend too much time preventing small blackouts. She argues that this theory ignores the fundamental fact that large blackouts start out as small problems that do not by themselves necessarily lead to blackouts. They lead to blackouts when small problems are not corrected in time.

Professor Vijay Vittal, Iowa State University, corroborates that the impact of such severe system failures could be mitigated by several approaches such as use of corrective controls or performing more effective analysis closer to real time.

The effective way to minimize disturbance propagation is to truly understand the common causes and design the appropriate solutions. The system needs to be addressed as a whole, implementing various planning, operations, maintenance, and regulatory measures in a coordinated way.

A possibility to prevent propagation of the disturbance throughout interconnected grid, but not weaken the grid during normal operation, is to design the interconnected power system to allow for intentional separation into stable islands or interrupt small amounts of load only when the system experiences major disturbances. As operators may not be able to act fast enough to take into account all data related to the on-line state of the system, separation actions should be done automatically. System Integrity Protection Schemes (SIPS) are wide-area automatic schemes that are designed to detect abnormal system conditions and initiate pre-planned automatic and corrective actions based on system studies [10–13]. They are also referred to as Special Protection Schemes (SPS), or Remedial Action Schemes (RAS).

They detect abnormal wide-area system conditions and trigger automatic actions to restore acceptable system performance. The initiating factor in implementing significant number of SIPS in the western part of the USA (WECC) has been to better protect the system against multiple contingencies, particularly after the 1994 and 1996 blackouts [7]. Designing the grid with appropriate measures for voltage control and advance warning systems such as wide-area protection and control would allow for both strong interconnected grids during normal operation (to make system more reliable and secure) and creation of predetermined islands only when necessary.

In conclusion, instead of weakening the grid, power industry needs to address deregulation as one important aspect in understanding the underlying causes of system-wide outages. The bulk power system was often not originally designed to transfer large amounts of power between neighboring systems. Individual power systems were interconnected to improve electrical network reliability by enabling neighboring utilities to support each other during stressed conditions. In recent years, deregulation has imposed additional requirements of high transfers from new generation sources to the load areas. At the same time, public pressures and the “Not in My Backyard” sentiment make it difficult to site transmission lines or major local generation sources, especially in the more densely populated heavy load areas, making the system expansion very expensive and difficult. In addition, recent disturbances have, more and more, been accompanied by voltage stability problems.

In summary, following are the main reasons for disturbances to turn into blackouts:

- Inability to prevent sequential tripping due to overloads, power swings, and voltage fluctuations
- Inadequate or faulty EMS/SCADA system (incl. alarm burst) – need for reliable alarm filtering, proper visualization, and analysis tools
- Protection miss-operation or unnecessary actions
 - Incorrect settings, e.g., impedance-based protective devices tripping on overloads (NERC developed regulations after the 2003 blackout to address it)
 - Hidden failures: uncovered application design flows or HW failures

- Inadequate design, e.g., application of impedance-based protective devices without the out-of-step blocking
- Inability of operators to prevent further propagation of the fast-developing disturbance
- Lack of coordinated response during developing disturbances, e.g., joint procedures between ISOs to deal with the problem quickly and effectively
- Generators tripping too early – generator tripping to be coordinated with the rest of the system

Blackout Prevention

The alarming increase in the number of major blackouts requires exploring new frontiers in deployment of well-defined and coordinated overall plans (planning, operations, and maintenance). As analysis of recent disturbances reveals some common threads among them, the conclusion is that propagation can be arrested and impact of disturbances reduced if knowledge gained is properly utilized [1–4].

The best way to minimize wide-area power system disturbances is to understand the leading causes. Study of blackout history in the past decade shows that in each case the reliability standards have been violated in some form or fashion demonstrating the priority for more stringent compliance enforcement of the standards, necessity to invest wisely in new transmission facilities that expand the reliability parameters of the grid in the “right” areas, new grid monitoring technologies, and especially the tools that help us make it simpler to manage the day-to-day operation.

Electric reliability and efficiency are affected by four segments of the electricity value chain: generation, transmission, distribution, and end use. Satisfactory system performance requires investments in all these segments of the system. Increasing supply without improving transmission and distribution infrastructure in the right locations, for example, may actually lead to more serious reliability issues.

The retirement and replacement of transmission equipment at the end of its useful life will be another important remedy for increasing failure rates and potential outages in the future. Aside from the aging infrastructure concerns, the transmission grid must be upgraded and expanded to continue to meet the growing demands. For example, high-voltage power

electronic devices allow more precise and rapid switching to improve system control and to help increase the level of power transfer that can be accommodated by the existing grid. Distributed energy technologies, if properly applied, could also play a role in relieving power flow demands on the transmission networks. In conclusion, while the new investments shall certainly include some new transmission lines, it will also encompass power delivery technologies such as series capacitors, single-phase operation of transmission lines, FACTS, HVDC links, energy storage, superconducting materials, and micro-grids.

The legislative and governmental commitments to build infrastructure, compliance with reliability requirements, and state, regional, and interregional plans are required to achieve the sustainable grid. Measure such as computerized control and data acquisition, phase-shifting transformers, coordination mechanisms, and electronic data exchange between operators are also some alternatives that improve the capability of the existing infrastructure and allow for a more robust power exchange.

Furthermore, reliable power system performance requires a balance of many critical components such as adequate reserve real and reactive power margins, reliable real-time telemetry and status monitoring, real-time state estimation, properly set, maintained, coordinated, and tuned protection and control systems, etc. Academic, industry, and governmental initiatives are required to set and enforce the standards for voltage control and reactive power practice, to improve system modeling and validation process, to enhance operator-training curriculum, and to assure that operators are assigned responsibility to take actions to prevent disturbance propagation.

In terms of modeling, quantitative analysis is needed to validate models of generators, turbines, and the associated controls to match actual system oscillations and damping. Likewise, dynamic loading or stability impact on protection devices (designed to operate for faulted conditions, e.g., tree contact) should be considered, and routine protection coordination studies using accurate model should be regularly performed. There are also concerns associated with protection and control application and settings when short-term market conditions for power transfers stress the equipment resulting in a risk of equipment outage.

Tools that improve real-time system monitoring, evaluation, visibility, security, congestion tracking, control, visualization, and information sharing about grid conditions over a wide region will allow operators to manage the grid more reliably on a day-to-day basis as well as in emergencies. Poorly recognized dynamic constraints can unnecessarily narrow operating limits and endanger reliability. Real-time security analysis tools are becoming increasingly critical for daily operation to visualize critical stability boundaries and to determine stability operating limits based on actual conditions.

Finally, the recent large-scale generation trips and remedial action responses have provided some very good benchmarks for combined small and large-scale analysis. Reexamination of traditional planning, operating, system design, protection applications, and device settings will help improve system response to slow or limit the spread of cascading outages. The frequency and varying impact levels of worldwide blackouts have provided the power industry with opportunities and supporting information to:

- Study the complex power system phenomenon to minimize propagation for future system-wide events using accurate and user friendly tools
- Validate the system studies against actual power system performance and governor modeling data
- Environmental and political factors limiting new addition of generation and transmission capabilities. Highlight the needed support for regulatory measures to ease wise grid expansions, grid reinforcements, and well-established and measured reliability enforcement process
- Operating capacity reserves and margins for transmission flows must remain available to allow system adjustments during unintended multiple contingency conditions. Enforce reliability requirements, for example, Planning Standards for Normal and Emergency Conditions
- Visit existing operating practices and real-time data exchange policies among control areas
- Make use of enhance maintenance practices and asset management tools
- Timely deploy Special Integrity Protection Schemes (SIPSs) to prevent spreading of the disturbance

Furthermore, protection systems are usually involved in major wide-area disturbances, sometimes

preventing further propagation, and sometimes contributing to the spread of the disturbances [14]. A very important lesson is that the design and operation of conventional protection and control schemes have to be scrutinized assuming stressed conditions. Improvements to existing protection systems can help prevent cascading and minimize the impact and number of wide-area disturbances. In general, a protection system should operate only for its designed conditions. Preventing further disturbance propagation should be achieved by designing and setting protection, and control schemes do not misoperate during major disturbance conditions. However, some experiences include relay operations that prevented further cascading by tripping on system conditions for which they were not designed.

Some key opportunities for improvement include coordinated adaptive protection and control systems and wide-area monitoring with advance warning systems, as elements of WAMPAC. The advance technology today promotes the concept of the “smart grid” – an integrated, electronically controlled power system that will offer unprecedented flexibility and functionality, and improve system reliability. The concept of the smart power delivery system includes automated capabilities to recognize problems, find solutions, and optimize the performance of the system.

In summary, corrective and preventive actions to avoid wide-area blackouts are the following:

- Provide operator with:
 - Tools to measure, monitor, assess, and predict both system performance and the performance of market participants
 - Training, incl. coordinated approach among control areas and use of dispatch training simulators
- Improve monitoring and diagnostics and control center performance
 - Availability of critical functions needs to be 99.99%
 - Assure adequate exchange of information between neighboring control centers
 - Secure real-time operating limits on a daily basis
 - Shared and consistent rating information required
- Use dynamic line ratings ambient temperature, wind, pre-contingency loading, etc., to better understand what the actual system margins are

- Better utilize available and standby generation in the area; implement black-start capabilities
- Improved maintenance and condition assessment of aging infrastructure
- Improve vegetation management
- Accurate generator, dynamic load, and reactive support device models, including renewable generation models
- Advanced algorithms and programs to assist the operator, such as “faster than real-time simulations”
- Protection coordination studies across regions and in coordination with equipment control and protection
 - Study and review protection designs on a regular basis, as system conditions change
 - Assure planned relay operation (e.g., distance relays not to trip on out-of-step and overload; ground overcurrent relays not to trip on high unbalance during high load, coordinated zone 3 operation, etc.)
 - Avoid hidden failures by adequate testing of not only individual relays but also overall relay applications
 - Increase the security of protection design in the areas vulnerable to blackouts
- Deploy Wide Area Monitoring, Protection, and Control (WAMPAC) systems to improve grid visibility and initiate corresponding actions

In conclusion, it is important to take a balanced approach to fixing the system as a whole by implementing various planning, operations, and maintenance measures and weighing the costs, performance impacts, and risks associated with each measure. One needs to evaluate how to operate and maintain the power system for the years to come to meet defined reliability objectives. There is no silver bullet solution to preventing blackouts, but there are general measures than can and should be taken to minimize impact of wide-area disturbances.

Each entity needs to focus on further process improvement, standardization, and better asset utilization, all parts of overall asset management strategy. This is the key to increased reliability and to protecting investments. Prudent capital investment in power system infrastructure has to be based on stringent

cost-benefit analysis to optimize investments. For example, increase in generation (“conventional” or renewable energy) has to be planned in conjunction with strengthening the transmission grid. Holistic, multi-year system planning is required to achieve efficient, reliable, and cost-effective grid. In addition, independent certification of technical systems and business processes can be an important element of assuring that proper actions are taken, processes implemented, and investments made.

System Integrity Protection Schemes

Examples of large blackouts in the past decade have shown that the risk of large blackouts is no longer acceptable and can lead to very large and unexpected social and financial consequences. Reduction of the risk of large system-wide disturbances and blackouts requires that system protection function be approached with the assistance of modern technologies in support of preserving system integrity under adverse conditions.

These schemes, defined as System Integrity Protection Schemes (SIPS) [10–13], are installed to protect the integrity of the power system or strategic portions thereof, as opposed to conventional protection systems that are dedicated to a specific power system element. The SIPS encompasses Special Protection System (SPS), Remedial Action Schemes (RAS), as well as other system integrity schemes such as Underfrequency, Undervoltage, Out-of-Step, etc. These schemes provide reasonable countermeasures to slow and/or stop cascading outages caused by extreme contingencies.

SIPS goal is to prevent propagation of disturbances for severe system emergencies caused by unplanned operating conditions and ensure system security. They stabilize the power system for equipment outages, N-2 (two key elements out of service) or beyond by:

- Preventing cascading overloading of the lines and transformers
- Arresting voltage decline
- Initiating pre-planned separation of the power system

Advanced detection and control strategies through the concept of SIPS offer a cohesive management of the

disturbances. With the increased availability of advanced computer, communication, and measurement technologies, more “intelligent” equipment can be used at the local level to improve the overall response. Traditional dependant contingency/event-based systems could be enhanced to include power system response based algorithms with proper local supervisions for security.

The IEEE Power System Relaying Committee has developed a worldwide survey on SIPS [12]. Figure 3 shows a summary of the overall SIPS purpose classification. The numbers of SIPS performing similar types of functions have been grouped to indicate the total number of SIPS types. For each type of SIPS scheme, the number of schemes serving a similar purpose has been indicated, with the following classification:

1. Essential: Prevent cascading outages
2. Increased Security: Minimize area affected by undesirable conditions
3. Increased Power Flow Capability: To extend transmission system rating without adding new transmission facilities or to delay enhancement of transmission networks
4. Important: Avoid difficult operating conditions
5. Normal: A better functioning of the network

Note that almost all classifications are evenly distributed (with exception of “Important” which is at 8%). The approximate even distribution of classifications of SIPS highlights the important role of SIPS in grid reliability and how SIPS are integrated part of the grid development worldwide.

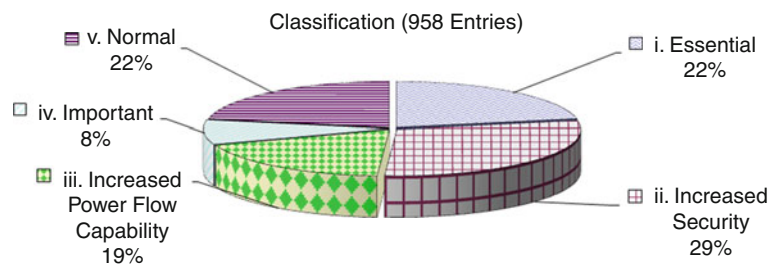
It is clear from Fig. 3 that the application of SIPS has become a component of a comprehensive total grid operation and protection philosophy. The fact that 22% of the entries are applications to address “normal” system conditions demonstrates that SIPS are no longer applied solely for system security purposes. In fact, close examination of Fig. 3 reveals SIPS applications can be viewed as two major categories:

- *Operational system improvement* (49% with three components: 19% Increased Power Flow, 8% Important, plus 22% Normal)
- *System security* (51% with two components, 22% Essential plus 29% for Increased Security) which at one time was the primary intent of SIPS

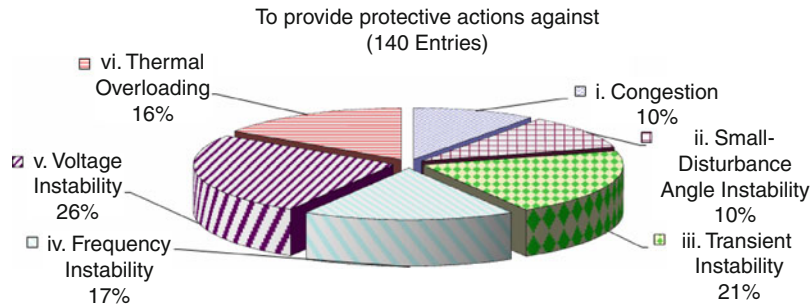
Figure 4 shows the intent of the various types of SIPS. The fact that voltage instability is the most often addressed phenomenon confirms that systems are now more complex than ever. The voltage stability phenomenon was firstly discovered at the beginning of 1980s as systems were getting more complex. The information in Fig. 4 correlates with the classifications in Fig. 3, demonstrating that worldwide SIPS are integrated components of various aspects of grid operation.

System Restoration

Another critical step in minimizing the impact of widespread blackouts is the need for effective and fast power system restoration. Returning equipment to service followed by quick restoration of power to the users is of paramount importance and can significantly minimize consequences of further outages.



Transmission Blackouts: Risk, Causes, and Mitigation. Figure 3
SIPS classification



Transmission Blackouts: Risk, Causes, and Mitigation. Figure 4
SIPS purpose

Today's technology can be used to our advantage for intelligent restoration. Some of the key elements for responsive restoration are the following:

- Well-defined procedures that require overall coordination within the restoring area, as well as with the neighboring electric networks.
- Reliable and efficient restoration software tools significantly aid operators and area coordinators to execute operating procedures and to make proper decisions. This tool is a part of EMS/SCADA system that provides voltage, frequency, excitation, outage status, and other data.
- Regular training sessions to assure effectiveness of the process. These sessions should include practice drill scenarios. The drill scenarios should incorporate any regional reliability or governmental policy requirements. For example, there may be a time delay requirement for load restoration after bulk power system has returned to service, to allow the system to stabilize. There may also be critical loads, which must be given higher priority in restoration.
- Substations need to be manned to open breakers or switches to clear re-energization pathways and establish ability to control load restoration.

Today's technology allows us to propel in designing schemes to aid in quick restoration. Even if advanced tools and procedures are in place to speed up restoration, there are limits on how fast the system can be restored depending on the type and distribution of generation. After the August 14, 2003 blackout in North America, it took considerable time to restore generation. Some of the units did not have capabilities to be put in service immediately (black-start

capabilities), and some units required longer time to be put on-line with full power (e.g., nuclear units due to security and steam turbines due to allowable ramp-up rates). Also of equal consideration is the type of load served, the system configuration, and the effects of connecting the load back to the network (Cold Load Pickup or Hot Load Pickup affects end user restoration time). While most of the cities have been restored in 5–9 h during the Italian blackout in September 2003, it took over a day to restore power back to Detroit and New York. Göran Andersson points out that in the recent Swedish-Danish blackout, the 400 kV grid was restored within 2 h, most customers were connected within 4 h, and the last customer reconnected within 6 h.

Mayer Sasson explains the slower pace for load restoration experienced in New York. The low-voltage network loads in New York City and Manhattan area are a highly meshed network system that affords a very high degree of reliability against localized outages. However, under blackout conditions, when a network is to be restored, the network is isolated into 100–200 MW portions that need to be re-energized at a time, requiring a time-consuming and careful process such that the inrush does not provide a set back to the restoration effort.

As discussed before, by designing the power system to transfer power across large distances and not providing enough reactive power close to the load or building the accompanying transmission lines may have detrimental effects on power system operation. Similarly, designing the power system not considering the effects on restoration efforts may have detrimental effects on the speed of restoration. In conclusion,

restoration time and system security could be significantly improved by planning of the generation mix and location considering not only market factors but incorporating value of reliable operation and faster restoration in the financial model. This approach would result in optimal, long-term investment strategies.

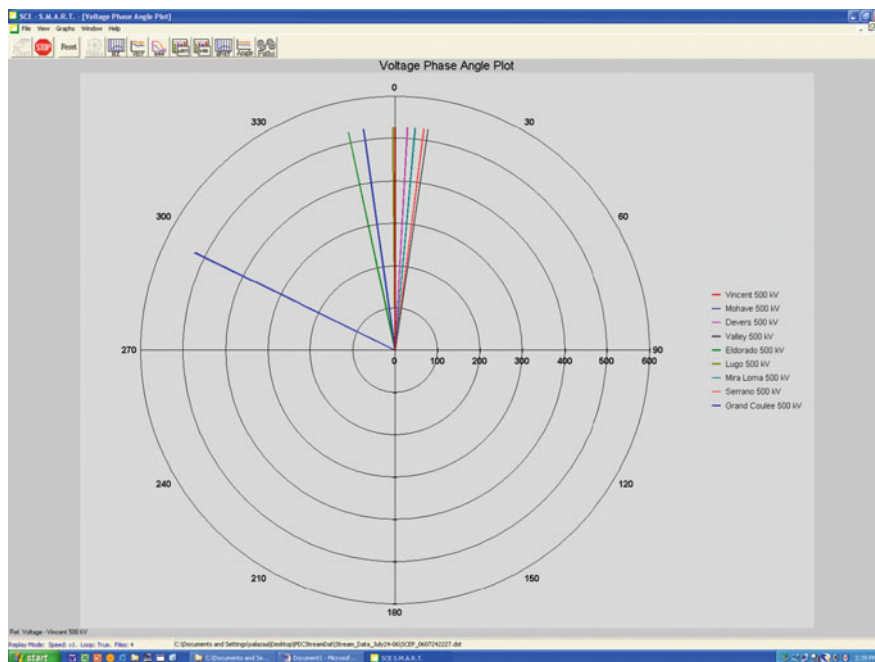
Future Directions

As the power grids worldwide have become more complex and are operated closer to the operating limits, applications of Wide Area Monitoring, Protection, and Control (WAMPAC) systems have become necessary to better manage the grid reliability and performance security [15]. Control areas within the interconnected grids or among countries are recognizing that bulk power systems should be treated as one system and power system professionals are challenged to upgrade the system with technologies that makes the task of grid control manageable. New technologies such as synchronized measurements have advanced to support commercial WAMPAC deployment so that implementation of

various applications is both possible and warranted, representing prudent investment.

Although large-scale demonstrations of using this technology are reported around the world, full-fledged productized systems have just started to be deployed. In the USA, those initiatives have been driven by transmission “smart grid” investments supported by DOE stimulus grants and NERC. WAMPAC systems using synchronized measurements have some unique deployment challenges as they require engagement of multiple users with diverse requirements and varying needs. Such large-scale systems not only offer a lot of promising reliability and financial benefits but also push the boundaries of conventional grid operations.

Synchronized measurement applications, using GPS synchronized Phasor Measurement Units (PMUs), offer large reliability and financial benefits for customers/society and the electrical grid when implemented across the interconnected grid. As measurements are reported 20–60 times per second, PMUs are well suited to track grid dynamics in real time. Compared to current EMS monitoring tools that use



Transmission Blackouts: Risk, Causes, and Mitigation. Figure 5
Key system voltage angles – WECC disturbance on July 24, 2006

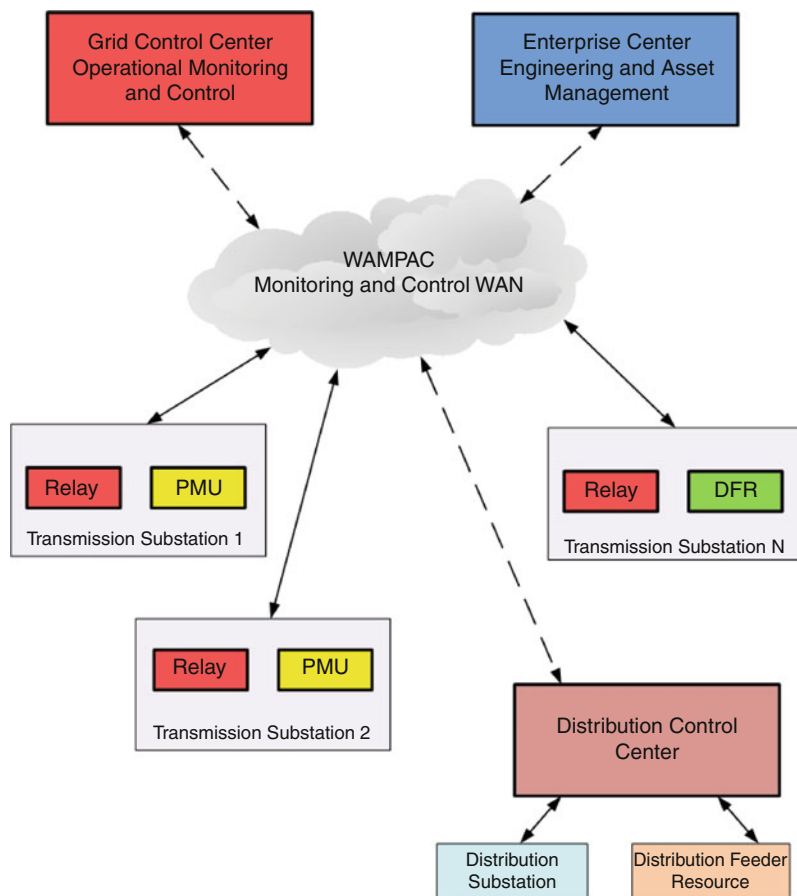
information from state estimation and SCADA over several second intervals, time-synchronized PMUs introduce the possibility of directly measuring the system state instead of estimating it based on system models and telemetry data. If implemented properly, the technology also allows for providing data integrity validation without added cost. This technology is instrumental for:

- Improving WAMPAC in real time including applications such as early warning systems, SIPS, detection and analysis of system stability, etc., and enabling faster system restoration
- Faster and more accurate analysis of vast number of data during transient events
- Validation and development of power system models

All of the above helps with avoidance and analysis of outages that may have extreme manifestation in blackouts. Due to its accuracy and wide-area coverage, synchronized measurement technology is a paradigm shift enabling unique tracking of power system dynamics. Synchronized measurement applications enable true early warning systems to detect conditions that lead to catastrophic events, help with restoration, and improve the quality of data for event analysis.

The display shown in Fig. 5 is a dynamic angle display during the disturbance in the western part of the USA in 2006. Those types of displays are used both for post-disturbance analysis and as an operational tool to help during the disturbance.

The conceptual WAMPAC system has interesting parallels with the human nervous system. Figure 6 shows the simplest concept of a WAMPAC system that



Transmission Blackouts: Risk, Causes, and Mitigation. Figure 6
Basic conceptual WAMPAC system

spans a complete utility transmission system, or even an entire unified operating region. This conceptual description focuses on likely future implementations, without concern of integrating today's persistent but obsolescent components like independent SCADA RTUs. WAMPAC is in fact ultimately capable of absorbing all of the functions of today's SCADA and EMS, including not only measurement and control but also state estimation and real-time contingency analysis.

In summary, Wide Area Monitoring, Protection, and Control (WAMPAC) is a must for transmission smart grid as it is necessary to increase probability to prevent blackouts by improving following applications [15]:

- Data analysis and visualization – Significant benefits are achieved with the ability to analyze events much faster in a synchronized fashion on-line or post-mortem.
- System reliability improvement by preventing cascading outages due to voltage, angular, and frequency instability, thermal overloads, and low frequency oscillations – These applications result in huge societal benefits.
- System operations and planning, improved modeling – Synchronized measurement enables paradigm shift with high reporting rates not available with any other technology and in developing accurate power system models.
- Market operations: Congestion Management and Locational Marginal Pricing – These applications enable large financial benefits by better grid utilization.

With today's technology, it is possible to tie all the monitoring, control, and protection devices together through an information network. The key to a successful solution is fast detection, fast and powerful control devices, communication system, and smart algorithms.

Conclusions

Power system is very complex and human-made. Our industry needs to keep planning, operating, and maintaining it as efficient and as reliable as possible, including preventing blackouts. There is a general understanding of blackouts caused by natural disasters (earthquake, hurricanes, etc.). However, system-wide

outages created by humans and/or not arrested due to suboptimal design should be easier to prevent. Analysis of large disturbances reveals some common threads among them, leading to the following conclusions:

- Need to understand the symptoms and root causes of the major disturbances and learn from the past blackouts.
- The power grid should not be operated under system conditions that have not been studied.
- Implement specific solutions to reduce the likelihood and propagation of outages.
- Restoration time could be reduced.

In summary, although it is not possible to avoid multiple contingency initiated blackouts, the probability, size, and impact of wide-area blackouts could be reduced and the propagation stopped.

Although electrical industry (utilities, ISOs, generation owners, regulators, universities, etc.) takes major initiatives after blackouts (such as investments in "smart grid" technologies), grids around the world always face host of new and old challenges. Applying a patchwork of individual measures to manage the grid is not sufficient. It is necessary to take a balanced approach to fixing the system as a whole by implementing a well-defined and coordinated overall strategy including various planning, operations, maintenance, and regulatory measures, weighing the costs, performance, and risks associated with each measure. It is important to always envision how power system should operate 10–30 years in the future; then design and overhaul it with this forward-looking approach.

Within the context of this approach, specific solutions to reduce the likelihood of outages can be addressed – because once the overall causes of wide-area disturbances are minimized, the smaller contributing factors are easier to handle, further diminishing the incidence of failures. The advent of advancements in information technology (IT), innovations in power system monitoring, and deployment of advance warning systems enables tools to arrest the grid from wide-area blackouts and meet the expectations for reliable power delivery. For example, while investing in strengthening the electrical grid infrastructure, such as rebuilding T&D grid and installing new generation and controls (e.g., reactive power devices, FACTS, HVDC), could not be replaced, "smart grid" WAMPAC

deployment is necessary and cost-effective way to improve grid reliability. Under normal conditions, and with sufficient automatic supports, operators are able to adequately control power system operation. However, the speed and complexity of disturbance phenomena makes control of the grid more suited to automated WAMPAC systems to respond to fast developing and complex disturbances.

Acknowledgment

The author expresses appreciation to the following Distinguished Industry Experts for their contributions and guidance: Göran Andersson, Miroslav Begovic, Daniel Karlsson, Bogdan Kasztenny, Mayer Sasson, Carlson Taylor, Chifong Thomas, Vijay Vittal, Jim Thorp, and Eric Udren. Special appreciation and recognition is extended to the Distinguished Expert Vahid Madani, whose ideas and work have been instrumental for the completeness of this chapter.

Bibliography

Primary Literature

- Novosel D, Begovic M, Madani V (2004) Shedding light on blackouts. *IEEE Power and Energy Magazine*, Jan/Feb 2004 (lead article)
- Madani V, Novosel D (2005) Getting a grip on the grid. *IEEE Spectrum*, pp 42–47 December 2005. Also, Taming the Power Grid in www.spectrum.ieee.org
- Novosel D, Madani V (2006) Blackout Prevention. In: *Yearbook of Science and Technology*, McGraw Hill, pp 34–38
- Madani V, Novosel D, Apostolov A, Corsi S (2004) Innovative solutions for preventing wide area disturbance propagation. In: *IREP symposium*, Cortina d'Ampezzo, Italy, Aug 2004
- WECC (1996) Western system coordinating council disturbance report for the power system outage that occurred on the western interconnection. www.NERC.com
- U.S.-Canada Power System Outage Task Force (2004) Final Report on the August 14, 2003 Blackout in the United States and Canada: Causes and Recommendations. www.FERC.gov, April 2004
- UCTE (2007) System disturbance on 4 Nov 2006. <http://www.ucte.com>
- Fairley P (2004) Unruly power grid. *IEEE Spectrum* 41(8):22–27
- Horowitz SH, Phadke AG (2003) Boosting immunity to blackouts. *IEEE Power and Energy Magazine*, Sept/Oct 2003
- CIGRE WG C2.02.24 (2007) Defense plan against extreme contingencies. *ELECTRA*, pp 46–61, Apr 2007
- Horowitz S, Novosel D, Madani V, Adamiak M (2008) System-wide protection. *IEEE Power and Energy Magazine*, Sept/Oct 2008
- Madani V, Novosel D, Begovic M, Adamiak M (2008) Application considerations in system integrity protection schemes (SIPS). *GE Protection and Control Magazine*, pp 25–30, May 2008
- Madani V, Novosel D, Horowitz S, Adamiak M, Amantegui J, Karlsson D, Imai S, Apostolov A (2010) IEEE PSRC report on global industry experiences with system integrity protection schemes (SIPS). *IEEE Trans PWRD* 25(4):2143–2155
- Novosel D, Bartok G, Henneberg G, Mysore P, Tziouvaras D, Ward S (2010) IEEE PSRC report on performance of relaying during wide-area stressed conditions. *IEEE Trans PWRD* 25(1):3–16
- Novosel D, Madani V, Bhargava B, Vu K, Cole J (2008) Dawn of the grid synchronization. *IEEE Power and Energy Magazine*, vol 6, Jan/Feb 2008, pp 49–60

Books and Reviews

- Kundur P (1994) *Power system stability and control*. McGraw-Hill, New York
- Massoud Amin S, Wollenberg BF (2005) Toward a smart grid. *IEEE Power and Energy Magazine*, Sept–Oct 2005
- Phadke AG, Thorp JS (2008) *Synchronized phasor measurements and their applications*. Springer, New York
- Taylor CW (1999) Improving grid behavior. *IEEE Spectrum* 36:40–45

Transport and Fate of Chemicals in the Environment, Introduction

JOHN S. GULLIVER

St. Anthony Falls Laboratory, Department of Civil Engineering, University of Minnesota, Minneapolis, MN, USA

Article Outline

Glossary
Introduction
Transport Processes
Chemical Fate
The Importance of Mixing
Resistance to Transport
Conclusion
Future Directions
Bibliography

Glossary

Concentration The quantity of a compound or chemical per unit volume, unit mass, or unit mole, where

1 mole = 6.02×10^{23} molecules of the chemical or compound. In this text, concentration in mass or moles per volume of water, mass per mass of solid, and moles per mole of gas will be discussed, depending upon the media of interest.

Convection The movement of a constituent with the movement of the fluid.

Density Total mass per unit volume.

Diffusion The spreading of fluid constituents through the motion inherent to atoms and molecules.

Diffusion coefficient A coefficient that describes the tendency of molecules to spread a constituent mass.

Dilution The mixing of a more concentrated solution with one that is less concentrated. The adage “The solution to pollution is dilution” is still used, sometimes appropriately, for many pollution and mitigation processes.

Kinematic viscosity The fluid viscosity divided by the fluid density, resulting in units that are similar to a diffusion coefficient, or length squared per time.

Turbulent diffusion The mixing of chemicals by turbulence, such that a turbulent diffusion coefficient can be defined separately from the temporal mean convection.

Introduction

Estimating the fate and transport of chemicals released into the environment is an interesting and challenging task. The environment can rarely be approximated as well mixed, and the chemicals in the environment often are not close to equilibrium. Thus, chemical fate and transport in the environment requires a background in the physics of fluid flow and transport, chemical thermodynamics, chemical kinetics, and the biology that interacts with all of these processes. The goal is to follow chemicals as they move, diffuse, and disperse through the environment. These chemicals will inevitably react to form other chemicals, in a manner that approaches, but rarely achieves, a local equilibrium. Many times, these reactions are biologically mediated, with a rate of reaction that more closely relates to an organism being hungry, or not hungry, than to first-order kinetics.

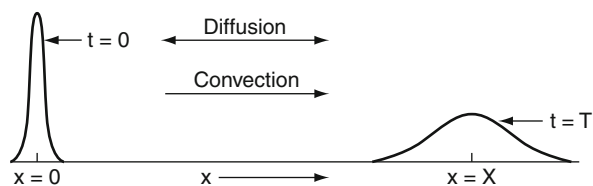
The global environment is large, on the chemical fate and transport scale. Individuals attempt to apply the mathematics of diffusion to the atmosphere, lakes,

ivers, groundwater, and the ocean, depending on the system for which the material is most applicable, and to transfer between these systems. Volatilization of a compound from a water body, condensation of a compound from the air, and adsorption of a compound from a fluid onto a solid are all interfacial transport processes. Thus, the fate and transport of chemicals in the environmental media of earth, water, and the atmosphere will be the topic. Contributions in this section will first attempt to formulate fate and transport problems such that they can be solved, regardless of the media or of the transport process, through the mathematics of diffusion (see chapters on ► [Chemicals in the Environment, Diffusive Transport](#); ► [Chemicals in the Environment, Dispersive Transport](#); ► [Chemicals in the Environment, Turbulent Transport](#); ► [Toxic Organic Chemicals](#); ► [Transport in the Environment](#)). Applications of these principals to specific media and between media will then be described (see chapters on ► [Transport with Jets and Plumes of Chemicals in the Environment](#); ► [Atmosphere-Water Exchange](#); ► [Sediment-Water Interfaces, Chemical Flux at](#); ► [River Fate and Transport](#); ► [Lake and Reservoir Fate and Transport of Chemicals](#); ► [Oceanic Fate and Transport of Chemicals](#); and ► [Subsurface Fate and Transport of Chemicals](#)).

Transport Processes

A transport process, as used herein, is one that moves chemicals and other properties of the fluid through the environment. *Diffusion* of chemicals is one transport process, which is always present. It is a spreading process, which cannot be reversed (without the involvement of another media such as in reverse osmosis). *Convection or advection* is the transport of chemicals from one place to another by fluid flow. The convection and diffusion of a chemical cloud, as represented in [Fig. 1](#), is the movement of the cloud and spreading of the cloud over time.

Turbulent diffusion is actually a form of advection, but the turbulent eddies tend to mix fluid with a random characteristic similar to that of the diffusion process, when viewed from enough distance. The representation given in [Fig. 1](#) could also be used to represent convection and turbulent diffusion, except that the pace of turbulent diffusion is normally more than one



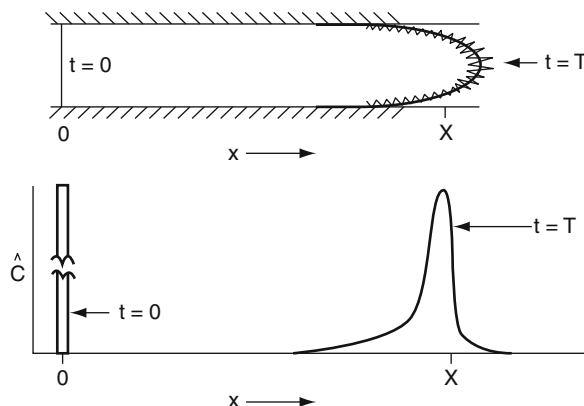
Transport and Fate of Chemicals in the Environment, Introduction. Figure 1

Illustration of convection and diffusion of a chemical cloud along the x-space coordinate (x-axis) [2]

order of magnitude greater than diffusion. This higher pace of turbulent diffusion means that diffusion and turbulent diffusion do not normally need to be considered together because they can be seen as parallel rate processes, and one has a much different time and distance scale than the other. If two parallel processes occur simultaneously, and one is much faster than the other, the second process can normally be ignored. This is discussed further in this section.

Dispersion is the combination of a nonuniform velocity profile and either diffusion or turbulent diffusion to spread the chemical longitudinally or laterally. Dispersion is something very different than either diffusion or turbulent diffusion because the velocity profile must be nonuniform for dispersion to occur. The longitudinal dispersion of a pipe flow is illustrated in Fig. 2. While there is diffusion of the chemical, the nonuniform velocity profile creates a dispersion that is much greater than would occur with diffusion alone. The other important difference is that *dispersion reflects the spreading of a cross-sectional mean concentration*, while diffusion represents the spreading of a local concentration. In some contexts, typically in atmospheric applications, turbulent diffusion is also considered to be a form of dispersion. This is only a semantic difference, and herein, the differentiation will continue to be between turbulent diffusion and the dispersion of a mean concentration.

Interfacial transfer is the transport of a chemical across an interface. The most studied form of interfacial transfer is absorption and volatilization, or condensation and evaporation, which is the transport of a chemical across the air–water interface. Another form of interfacial transfer would be adsorption and desorption, generally from water or air to the surface of a particle of soil, sediment, or dust. Illustration of both of these forms of interfacial transfer will be given later in this section.



Transport and Fate of Chemicals in the Environment, Introduction. Figure 2

Illustration of longitudinal dispersion of a tracer "plane" at $t = 0$ to a dispersed "cloud" at $t = T$. is the cross-sectional mean concentration [2]

Finally, there is *multiphase transport*, which is the transport of more than one phase, usually partially mixed in some fashion. The settling of particles in water or air, the fall of drops, and the rise of bubbles in water are all examples of multiphase transport. Figure 3 illustrates three flow fields that represent multiphase transport.

Mass transport problems are solved with the *diffusion equation*, often represented as

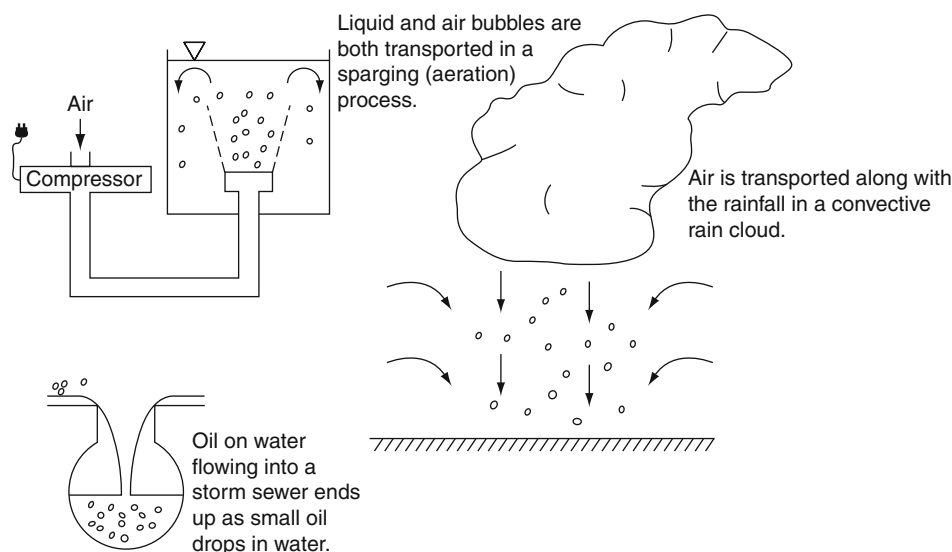
$$\frac{\partial C}{\partial t} + u \frac{\partial C}{\partial x} + v \frac{\partial C}{\partial y} + w \frac{\partial C}{\partial z} = D \left[\frac{\partial^2 C}{\partial x^2} + \frac{\partial^2 C}{\partial y^2} + \frac{\partial^2 C}{\partial z^2} \right] + S \quad (1)$$

1 ← 2 → | | ← 3 → | 4

where C is the concentration of a chemical; t is time; u , v , and w represent the temporal mean velocity in the x , y , and z directions, respectively; and D represents a diffusion coefficient. The first term (1) on the far left of Eq. 1 represents the rate of accumulation of chemical concentration. The second terms (2) represent the mean convection of the chemical. The third terms (3) to the right of the equal sign represent either diffusion or turbulent diffusion of chemical. The fourth term (4) represents the multitude of reactions that are possible in a fluid in environmental media.

Chemical Fate

Chemical fate is the eventual short-term or long-term disposition of chemicals, usually to another



Transport and Fate of Chemicals in the Environment, Introduction. Figure 3

Illustration of sparged multiphase transport. In these two cases, air bubbles create a water flow, and rain drops create an airflow. The oil drops and rain drops create an airflow. The oil drops do not have a significant rise or fall velocity in water and are simply transported [2]

chemical or storage. Some examples that fit the concept of short-term and long-term fate are given in Table 1. If a polychlorinated biphenyl (PCB) compound is in groundwater, the media is soil and water. The “short-term” fate will be that the PCB will primarily adsorb to the soil. The “long-term” fate is that the chemical will desorb, when the PCB-laden water has left, and eventually be bioremediated by microbacteria looking for carbon sources. If this PCB is in the atmosphere, it will be adsorbed primarily to aerosols and particles in the short-term, while its long-term fate will probably be photocatalyzed degradation.

There are as many examples of short-term and long-term fate as there are chemical-media combinations. An important consideration for this topic is whether one is interested in short-term or long-term fate. This is often a question to be answered by toxicologists because it is the most toxic forms of a chemical that are of most interest.

The Importance of Mixing

Mixing is a rate-related parameter in that most rates of reaction or transport are dependent upon mixing in

Transport and Fate of Chemicals in the Environment, Introduction. Table 1 Examples of short-term and long-term fate

Chemical	Media	Short-term fate	Long-term fate
PCB	Soil and water	Adsorbed to soil	Bioremediated degradation
PCB	Atmosphere	Adsorbed to aerosols	Photocatalyzed degradation
CO ₂	Water	Reactions to carbonate and bicarbonate	Photosynthesis to oxygen and biomass
Benzene	Water	Adsorbed to suspended particles	Bioremediated degradation
Ammonia	Soil and water	Reaction to ammonium	Bioremediated degradation to N ₂

environmental systems. When mixing is dominant (the slowest process), the first-order rate equation can be described as

$$\begin{aligned} \text{Rate of process} &= \text{Mixing parameter} \\ &\times \text{Difference from equilibrium} \end{aligned} \quad (2)$$

Thus, two items are needed to compute the rate of the process: the equilibrium concentrations for all species involved and the mixing rate parameter. A common example would be dissolved oxygen concentration in aquatic ecosystems.

One of the most common chemicals of concern in water bodies is oxygen. Without sufficient oxygen, the biota would be changed because the “desirable” organisms in the water body require oxygen to live. The rate of oxygen transfer between the atmosphere and a water body is therefore important to the health of the aquatic biota. For air-water oxygen transfer, Eq. 2 can be formulated as

$$\frac{dM}{dt} = K_L A \left(\frac{C_a}{H} - C \right) \quad (3)$$

where dM/dt is the rate of mass transfer into the water, K_L is a bulk oxygen transfer coefficient, A is the surface area for transfer, C_a is the concentration of oxygen in the air, H is a coefficient that partitions oxygen between the air and water at equilibrium (called Henry’s Law constant for liquids and gas equilibrium), and C is the concentration of oxygen in the water. Air is 20.8% oxygen, so the concentration of oxygen in the atmosphere is determined primarily by atmospheric pressure. Henry’s Law constant for oxygen is a function of pressure as well as temperature. Thus, the equilibrium concentration of oxygen is influenced by the thermodynamic variables pressure and temperature. The rate parameter is $K_L A$, which has units of volume/second. The difference from equilibrium partitioning is represented by $C_a/H - C$. It is C that is typically needed to bring as close to equilibrium with the atmosphere as possible, and the means to do it is by having a large dM/dt . This usually means a large $K_L A$ because it would be difficult to alter either C_a or H in the atmosphere. While the surface area is often established by the boundary conditions, K_L is determined by the turbulence and diffusion coefficient (i.e., mixing) close to the water surface and represents the rate of mixing per unit surface area. Thus, the primary variable that can be changed in order to increase dM/dt is the mixing parameter represented by K_L . Some further examples

of mixing rate and equilibrium parameters in environmental processes are given in Table 2.

Resistance to Transport

An important concept for environmental transport is *resistance*. That is, the inverse of a rate parameter is a resistance to chemical transport, or in equation form:

$$\frac{1}{\text{Rate parameter}} = \text{Resistance to chemical transport} = R \quad (4)$$

Figure 4 gives an example of the adsorption of a compound to suspended sediment, modeled as two resistances in series. At first, the compound is dissolved in water. For successful adsorption, the compound must be transported to the sorption sites on the surface of the sediment. The inverse of this transport rate can also be considered as a resistance to transport, R_1 . Then, the compound, upon reaching the surface of the suspended sediment, must find a sorption site for adsorption. This second rate parameter is more related to surface chemistry than to diffusive transport, and is considered as a second resistance, R_2 , that acts in series to the first resistance. The second resistance cannot occur without crossing the first resistance of transport to the sorption site, so they must occur in series.

Now, if R_1 is much greater than R_2 , it can be assumed that R_2 is zero without compromising the accuracy of the rate calculation. In electric circuits, two resistances applied in series are simply added together in calculating the line resistance. The same is true for resistance to chemical transport. If R_1 is 1,000 resistance units and R_2 is 1 resistance unit, R_2 can be ignored and still be within 99.9% of the correct answer.

Another example is the air–water transfer of a compound, illustrated in Fig. 5. This example will be used to explain volatile and nonvolatile compounds. There is resistance to transport on both sides of the interface, regardless of whether the compound is classified as volatile or nonvolatile. The resistance to transport in the liquid phase is given as $R_L = 1/K_L$. If chemical transfer is being described through an equation like Eq. 3, the resistance to transfer in the gas phase is given as $R_G = 1/(HK_G)$. The equilibrium constant is

Transport and Fate of Chemicals in the Environment, Introduction. Table 2 Examples of important mixing rate and equilibrium parameters in environmental processes

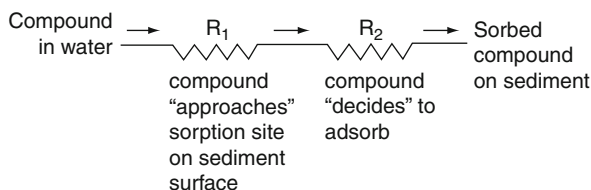
Process	Mixing-rate-related parameter	Equilibrium parameter
1. Treatment processes		
Coagulation/flocculation	Size of coagulation and flocculation basins and proper mixing (residence time)	Dose of coagulants (alum)
Softening	Design of softening tank to increase mixing	Dose of softening agent (lime)
Settling	Design of settling basin to reduce mixing	→ 0
Chlorination	Design of chlorination and dechlorination chambers for proper mixing and res. time	Dose of chlorine
Filtration	Size of filter bed	Length of time before backflushing
2. Surface waters		
Oxygen transfer	Diffusion and turbulent mixing	Atmospheric O ₂ conc. Henry's Law constant
Volatilization of pollutants	Diffusion and turbulent mixing	→ 0
Toxic spills	Diffusion and turbulent mixing	Spill–water equilibrium
Internal loading of nutrients	Hypolimnetic mixing	O ₂ concentration in hypolimnion
Sorption onto suspended sediments	Turbulent mixing exposes chemicals to sediment	Sediment–water partitioning
3. Atmosphere		
Greenhouse gases (CO ₂ , CH ₄)	Turbulent mixing	Atmospheric concentrations
Volatilization of spills	Turbulent mixing and dispersion	→ 0
Aerosols	Turbulent mixing, dispersion, and settling	None
4. Groundwater and sediments		
Spills	Advection and dispersion	→ 0
O ₂	Diffusion and advection	Atmospheric conc. of O ₂
Sorption of soil	Advection and diffusion	Equilibrium soil–water partitioning

in the R_G equation because the equivalent waterside concentrations are being used to represent the concentration difference from equilibrium, and the gas phase resistance needs to be a resistance to an equivalent water concentration.

The gas phase and the liquid phase resistances are applied in series. In general, gas film coefficients are roughly two orders of magnitude greater than liquid film coefficients. It is also true that Henry's Law

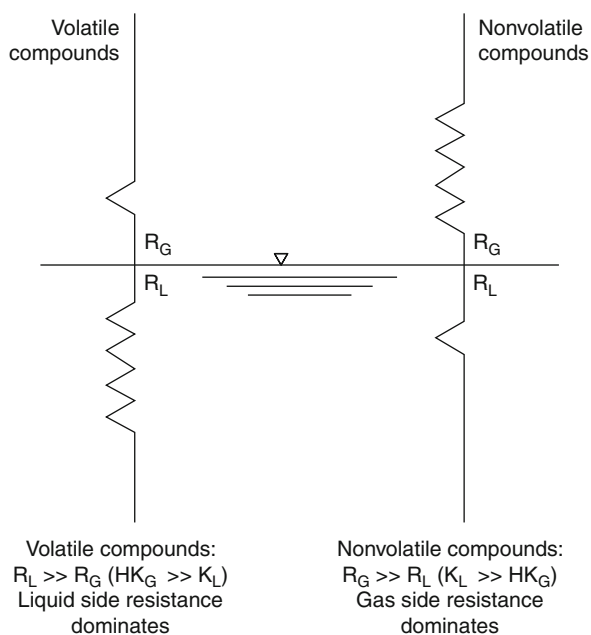
constant, H , varies over many orders of magnitude as the transported compounds are varied. Nitrogen gas, for example, has a Henry's Law constant of approximately 15, using mass concentrations. The herbicide, atrazine, has a Henry's Law constant of 3×10^{-6} . Thus, the ratio R_G/R_L would vary by seven orders of magnitude between nitrogen gas and atrazine.

If these orders of magnitude are put into a series resistance equation,



Transport and Fate of Chemicals in the Environment, Introduction. Figure 4

Adsorption analogy to two resistors in a series: adsorption of an organic compound to sediment [2]



Transport and Fate of Chemicals in the Environment, Introduction. Figure 5

Air-water transfer analogy to two resistors in a series [2]

$$R = R_L + R_G = \frac{1}{K_L} + \frac{1}{HK_G} \quad (5)$$

Due to the Henry's Law constants, it can be seen that for nitrogen gas, $R \cong R_L$, and for atrazine, $R \cong R_G$. If a typical ratio of $K_G/K_L \sim 100$ is applied, $R_G = R_L$ when $H = 0.01$.

Now, the mass transfer between phases is given as

$$\frac{dM}{dt} = \frac{A}{R} \left(\frac{C_a}{H} - C \right) \quad (6)$$

or

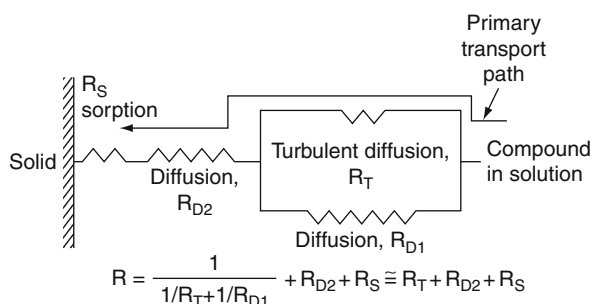
$$\frac{dM}{dt} = \frac{A}{1/K_L + 1/HK_G} \left(\frac{C_a}{H} - C \right) \quad (7)$$

Nitrogen gas would be a volatile compound because the equilibrium is strongly to the gas phase, and there is little gas phase resistance to its transfer, i.e., $1/K_L \gg 1/(HK_G)$. For that reason, N_2 is generally called a gas, as are many other volatile compounds such as methane, oxygen, and propane.

Atrazine, on the other hand, would be a nonvolatile compound, $1/(HK_G) \gg 1/K_L$, because equilibrium is strongly to the liquid phase due to the small Henry's Law constant. There is also a strong gas phase resistance to the transfer. Atrazine was manufactured to remain in the liquid phase, where it will act as a herbicide, rather than in the gas phase, where the farm personnel will be breathing this toxic chemical. If you were going to pick a nonionic compound that is not made by humans from the list of those that are gas or liquid in our environment, a good guess is that it would be a volatile or semivolatile compound. There are only a few environmental compounds that are nonvolatile. Remarkably, one of them is water. While the atmosphere may be as much as 3% water, the water bodies in the world are very close to 100% water. The equilibrium is strongly to the liquid side, due to a large equilibrium-partitioning constant.

One theme of this discussion can now be stated as follows: when transport processes occur in series, it is the slower transport processes that are important for chemical transport calculations because the *resistance* to transport is large, just as the large resistors of a series in an electronic circuit are the most important.

Now is the time for the second theme: when transport processes occur in parallel, the fast transport process with the low resistance dominates. The result is the opposite of resistances in series. Figure 6 illustrates this concept with the transport of a compound from a water body to a sorption site on a solid. In the bulk solution, there is diffusion and turbulent diffusion occurring simultaneously. Transport can occur due to either process, so there are two different paths that may be followed, without the need of the other path. These transport processes are operating in parallel, and the faster transport process will transport most of the



Transport and Fate of Chemicals in the Environment, Introduction. Figure 6

Transport to a sorption site and the resistor analogy [2]

compound. The analogy to electronic circuits applies in this case as well. Beginning with a compound in solution in Fig. 6, there are two parallel transport paths, each with a resistance to transfer. Most of the compound will be transported through the path with the least resistance. Many times, the path with the greater resistance can be ignored because the quantity of compound transported through this path is very small. When the compound comes close to the solid, however, the turbulent diffusion dissipates because eddies become so small that they are dissipated by the viscous action of the water. Now, one is back to one transport path, with the act of sorption and diffusion acting in series. Thus, the slowest transport path once again becomes the important process.

The overall resistance to the sorption process illustrated in Fig. 6 can be written as follows:

$$R = \frac{1}{1/R_T + 1/R_{D1}} + R_{D2} + R_S \cong R_T + R_{D2} + R_S \quad (8)$$

where R_T , R_{D1} , R_{D2} , and R_S are the resistances to turbulent transport, diffusive transport in the bulk of the fluid, diffusive transport near the solid surface, and adsorption, respectively. It can be seen that, in Fig. 6 and Eq. 8, the resistance due to diffusion in the bulk of the fluid can be neglected because turbulent diffusion is a parallel path. The resistance due to diffusion only needs to be considered when there is no parallel path for turbulent diffusion, such as very near the surface of the solid. Thus, R_{D1} can be ignored, but not R_{D2} .

Conclusion

In this chapter, some of the topics that will be covered and applied in the other chapters of the section have been introduced, where the physics of mass transport are essential. These and similar engineering concepts will be revisited throughout the first half of this volume, in an attempt to develop models in the environmental fate and transport of chemicals that are close to realistic, but can be solved, even if that solution is approximate.

Future Directions

Global climate change is a topic for the present and future because society has only begun to assess and tackle the causes, implications, and remediation of the new Anthropocene [1]. There are global warming gases (CO_2 and CH_4) and global cooling gases (refrigerants) that are transported between the oceans and the atmosphere and could be transported between the atmosphere and the earth (carbon sequestration). These global transport rates are not quick, so we humans will be dealing with transport to a greater extent in the future. On a more local scale, urbanization is creating nonsource pollution of water and the atmosphere that is increasing. Roads, parking lots, and rooftops are often directly connected to lakes and rivers through storm sewers, and the pollution of urban activities is not insubstantial. This will be a bigger concern as discovery about this source of pollution occurs. Finally, the nitrogen of various forms released from agricultural activities continues to create hypoxic (dead) zones when the rivers have transported this nutrient to the oceans. The nitrification–denitrification processes as organic nitrogen is decomposed to ammonia, utilized by aerobic bacteria to make nitrites and nitrates, and finally converted by anaerobic bacteria to nitrogen gas occurring in streams and rivers add an interesting complexity to the fate and transport calculations.

Bibliography

1. Ellis E (2011) A man-made world. The Economist London, UK, 26 May 2011
2. Gulliver JS (2007) Introduction to chemical transport in the environment. Cambridge University Press, Cambridge, UK

Transport in the Environment

JOHN S. GULLIVER

St. Anthony Falls Laboratory, Department of Civil Engineering, University of Minnesota, Minneapolis, MN, USA

Article Outline

Glossary

Definition of Transport in the Environment

Introduction

Development of the Diffusion Equation

Adsorption and Desorption in Sediments and Soils

Example Applications of the Diffusion Equation

Conclusion

Future Directions

Bibliography

Glossary

Adsorption The process of dissolved chemicals sticking to a solid.

Convection The movement of a constituent with movement of the fluid.

Desorption The detachment of a chemical from a solid.

Diffusion The spreading of fluid constituents through the motion inherent to atoms and molecules.

Diffusion coefficient A coefficient that describes the tendency of molecules to spread a constituent mass.

Dirac delta An impulse of a given quantity (mass) that occurs over an infinitely short time or space.

Kinematic viscosity The fluid viscosity divided by the fluid density, resulting in units that are similar to a diffusion coefficient, or length squared per time.

Laminar flow Flow that has no turbulent eddies, where the fluid flows in laminas and diffusion creates the mixing of the fluid.

Retardation factor A divisor that indicates the slowing of chemical movement through a media due to adsorption.

Reynolds number The ratio of inertial to viscous forces, resulting in a meaningful velocity times a meaningful distance divided by kinematic viscosity.

Definition of Transport in the Environment

In this section various solution techniques for the convection-diffusion equation are reviewed, which is generally defined as the mass transport equation with diffusive terms. These techniques will be applied to chemical transport solutions in sediments. There are also a number of applications to chemical transport in biofilms. There are many other applications of the convection-diffusion equation, but they require more background with regard to the physics of mixing processes, which will be addressed in later sections of the volume.

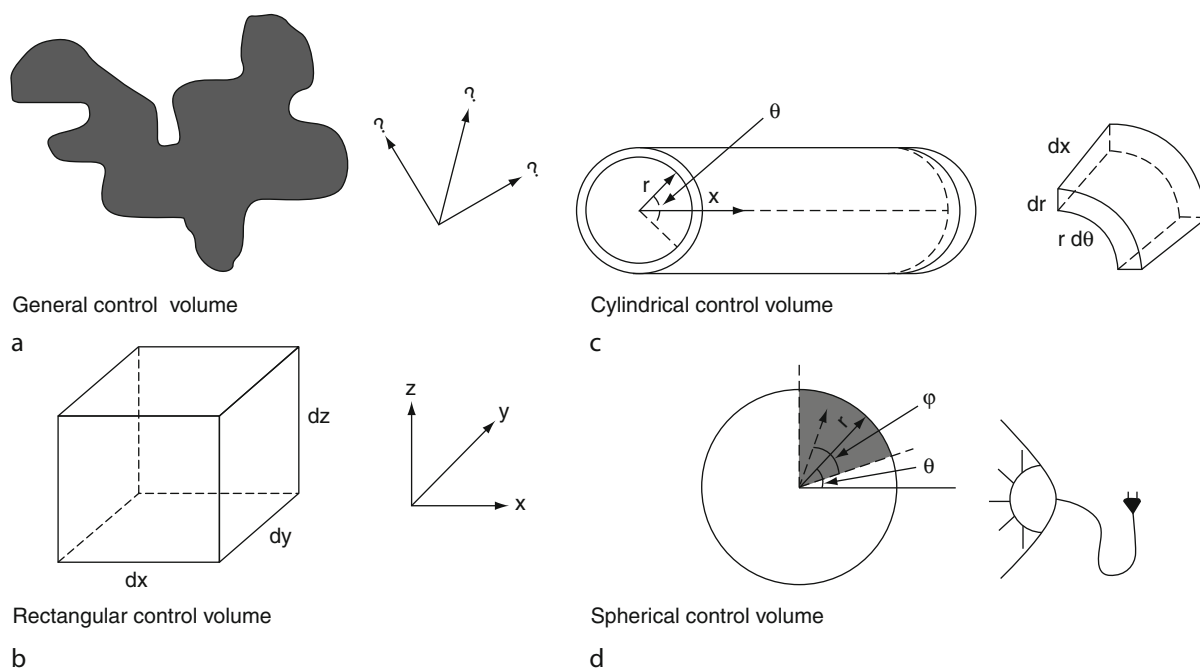
Introduction

What is *mass (or chemical) transport*? It is the transport of a solute (the dissolved chemical) in a solvent (everything else). The solute is the dissolvee and the solvent is the dissolver. There are liquids that are generally classified as solvents because they typically play that role in industry. Some examples would be degreasing and dry-cleaning solvents, such as trichloroethylene (TCE). In environmental applications, these “solvents” are the solutes, and water or air is usually the solvent. In fact, when neither water nor air are the solvents, a general term “nonaqueous phase liquid,” or NAPL, is applied. NAPL is defined as a liquid that is not water, which could be composed of any number of compounds.

The substance being transported can either be dissolved (part of the same phase as the solvent) or particulate substances. The diffusion equation will also be discussed by considering mass conservation in a fixed control volume. The mass conservation equation can be written as:

$$\text{Flux rate IN} - \text{Flux rate OUT} + \text{Rate of} \\ (\text{Sources} - \text{Sinks}) = \text{Rate of Accumulation} \quad (1)$$

Now that there is our mass conservation equation, it must be decided which control volume would be the most convenient for our applications. The control volumes used most for this type of mass balance are given in Fig. 1. The general control volume, given in Fig. 1a, is used for descriptive purposes, to maintain generality. It is rare that one works with something that approximates such a contorted control volume. The control volumes that are used in practice are given



Transport in the Environment. Figure 1

Common control volumes found in engineering texts and (for the latter three) used in solving the diffusion equation. (From [1])

in Figs. 1b, c, and d. For the environmental applications of chemical transport, the rectangular control volume, Fig. 1b, has proven to be the most useful. The cylindrical control volume, Fig. 1c, is used to make pipe or tube flow problems easier to solve, and the spherical control volume, Fig. 1d, is often helpful when dealing with transport in and around particles or drops. For this control volume, it is convenient to imagine a light being shined along the axis, which casts a shadow of the vector on to a plane normal to the light. The φ angle is measured from the reference axis to the shadow in this plane.

A rectangular control volume will be used for the development of our mass conservation (diffusion) equation.

Development of the Diffusion Equation

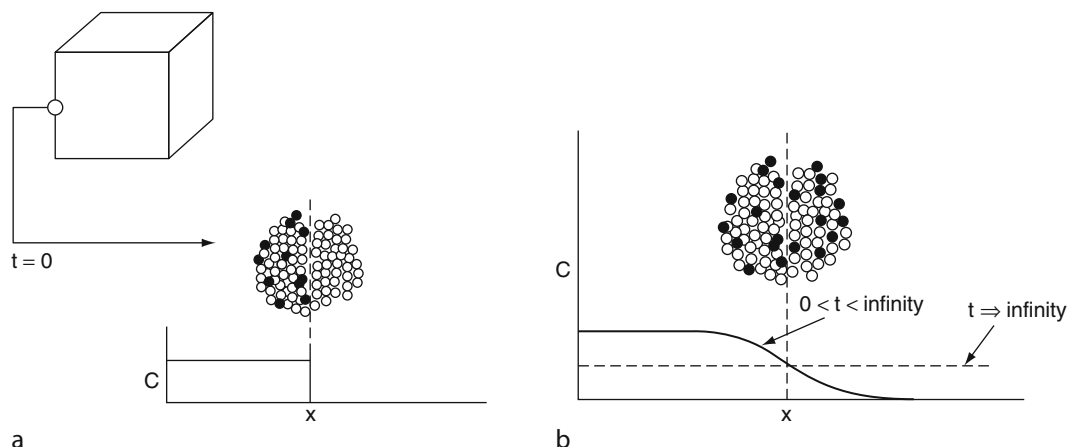
The diffusion equation will be developed by considering each term in Eq. 1 separately. In addition, the flux terms will be divided into diffusive and convective flux rates.

Diffusive Flux Rate

The molecules of a fluid “at rest” are still moving because of their internal energy. They are vibrating. In a solid, the molecules are held in a lattice. In a gas or liquid, they are not, so they move around because of this vibration. Since the molecules are vibrating in all directions, the movement appears to be random. Diffusive fluxes are described by Fick’s law [9], given in the section by Dr. Cussler on diffusion. For this purpose, let us consider one side of our control volume, normal to the x -axis, with an area A_x , shown in Fig. 2. Fick’s law describes the diffusive flux rate as:

$$\begin{aligned} \text{Diffusive flux rate (g/s)} \\ = -D(\text{m}^2/\text{s}) \frac{\partial C}{\partial x} (\text{g/m}^4) A_x(\text{m}^2) \end{aligned} \quad (2)$$

where C is concentration of the solute (tracer), D is the diffusion coefficient of the solute in the solvent (water), which relates to how fast how and far the tracer molecules are moving to and fro, and $\partial C/\partial x$ is the gradient of concentration with respect to x , or the slope of C



Transport in the Environment. Figure 2

Illustration of net diffusive flux through one side of the rectangular control volume. (From [1])

with x , as shown in Fig. 2. Thus, the diffusive flux rate depends upon the diffusion coefficient and the *gradient* of concentration with distance.

Convective Flux

The convective flux rate into our control volume is simply the chemical mass carried in by convection. If the same box of Fig. 2 is considered, except with a velocity component u in the x -direction, the convective flux rate into the box from the left-hand side is:

$$\begin{aligned} \text{Convective flux rate (g/s)} &= \text{Velocity component} \\ &\text{normal to surface (m/s)} \times \text{Surface area (m}^2\text{)} \\ &\times \text{Concentration (g/m}^3\text{)} \end{aligned} \quad (3)$$

or

$$\text{Convective flux rate} = uA_xC \quad (4)$$

where u is the component of velocity in the x -direction and A_x is the surface area normal to the x -axis on that side of the box. All six sides of our box would have a convective flux rate through them, just as they would have a diffusive flux.

Rate of Accumulation

The rate of accumulation is the change of chemical mass per unit time, or:

$$\text{Rate of accumulation (g/s)} = V(\text{m}^3) \frac{\partial C}{\partial t} (\text{g/m}^3/\text{s}) \quad (5)$$

where V is the volume of our box.

Source and Sink Rates

The solute chemical can appear or disappear through chemical reaction. In addition, interfacial transfer is often integrated over the control volume and considered as a source or sink throughout the control volume. This type of pseudo-reaction can be of significant help in solving chemical transport problems when averages over a larger control volume, such as cross-sectional mean concentrations, are being computed. For both cases (chemical reactions and pseudo-reactions), the source and sink rates are given as:

$$\text{Source} - \text{sink rate (g/s)} = S (\text{g/m}^3/\text{s}) V (\text{m}^3) \quad (6)$$

where S is the net source/sink rate per unit volume. The particular reactions that a given chemical is likely to undergo will determine the form of S used in Eq. 6. These are listed in Table 1. The source/sink term could be a combination of two or more of these reactions. For convenience in determining analytical solutions to the diffusion equation, most source/sink terms are approximated as either a first-order or zero-order reaction.

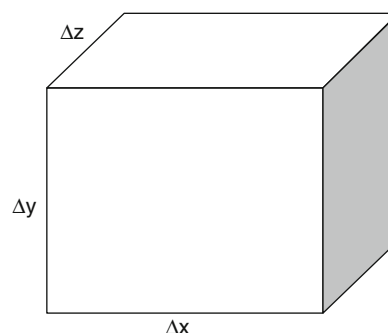
Transport in the Environment. Table 1 Common source and sink terms used in the convection-diffusion equation

Source/sink name	Equation	Units of constant
Zero order	$S = k_o$	$k_o - \text{g/m}^3\text{-s}$
First order	$S = k_1 C$	$k_1 = \text{s}^{-1}$
Second order	$S = k_2 C^2$	$k_2 - \text{m}^3/\text{g-s}$
Independent variable	$S = k_{1i} P^a$	$k_{1i} - \text{s}^{-1}$
	$S = k_{2i} PC^b$	$k_{2i} - \text{m}^3/\text{g-s}$
Monod kinetics ^c	$S = \frac{\mu_m C}{k_c + C} P$	$\mu_m = \text{maximum growth rate (s}^{-1}\text{)}$
		$k_c = \text{half-saturation coefficient (g/m}^3\text{)}$

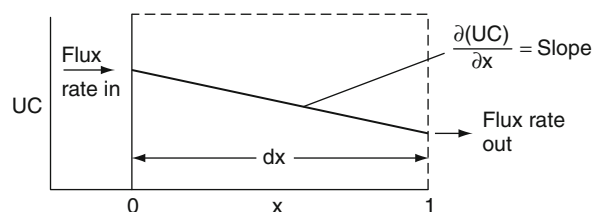
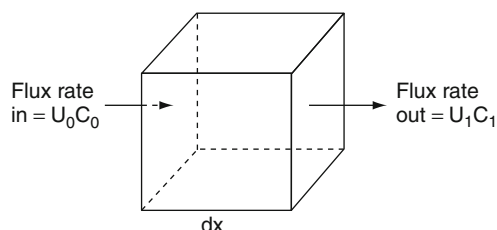
^aIf P is nearly constant, then k_{1i} can be provided as a zero-order term

^bOften called second order

^cCommon for biologically mediated reactions



Transport in the Environment. Figure 3
Dimension of the rectangular control volume



Transport in the Environment. Figure 4
Illustration of the x -component of mass flux rate into and out of the rectangular control volume. (From [1])

Mass Balance on Control Volume

A mass balance on one compound in our box is based upon the principle that whatever comes in must do one of three things: be accumulated in the box, flux out of another side, or react in the source/sink terms. If it seems simple, it is.

We will begin by assigning lengths to the sides of our box of dx , dy , and dz , as shown in Fig. 3. Then, for simplicity in this mass balance, we will arbitrarily designate the flux as positive in the $+x$ -direction, $+y$ -direction, and $+z$ -direction. The x -direction flux, so designated, is illustrated in Fig. 4. Then, the two flux terms in Eq. 1 become:

$$\text{Flux rate in} + \text{Difference in flux rate} = \text{Flux rate out} \quad (7)$$

or, because a *difference* can be equated to a *gradient times the distance over which the gradient is applied*:

$$\text{Flux rate out} - \text{Flux rate in} = \text{Gradient in flux rate} \times \text{Distance} \quad (8)$$

Equation 8 can thus be applied along each spatial component as:

$$\text{Flux rate(out - in)}_x = \frac{\partial}{\partial x} (\text{flux rate}) dx \quad (9a)$$

$$\text{Flux rate(out - in)}_y = \frac{\partial}{\partial y} (\text{flux rate}) dy \quad (9b)$$

$$\text{Flux rate(out - in)}_z = \frac{\partial}{\partial z} (\text{flux rate}) dz \quad (9c)$$

Convective flux rates. The convective and diffusive flux rates are dealt with separately. They will eventually be separated in the final diffusion equation, and it is convenient to make that break now. The x -component of the convective flux rate is equal to the x -component

of velocity times the concentration times the area of our box normal to the x -axis. Therefore, in terms of convective flux rates, Eq. 9a becomes:

$$\begin{aligned}\text{Convective flux rate(out - in)}_x &= \frac{\partial}{\partial x}(u C A_x) dx \\ &= \frac{\partial}{\partial x}(u C) dx dy dz\end{aligned}\quad (10a)$$

Because the normal area, $A_x = dy dz$, of our box does not change with x , it can be pulled out of the partial with respect to x . This is done in the second part of Eq. 10a. The same can be done with the y - and z -components of the convective flux rate:

$$\begin{aligned}\text{Convective flux rate(out - in)}_y &= \frac{\partial}{\partial y}(v C A_y) dy \\ &= \frac{\partial}{\partial y}(v C) dx dy dz\end{aligned}\quad (10b)$$

$$\begin{aligned}\text{Convective flux rate(out - in)}_z &= \frac{\partial}{\partial z}(w C A_z) dz \\ &= \frac{\partial}{\partial z}(w C) dx dy dz\end{aligned}\quad (10c)$$

Finally, adding Eqs. 10a, 10b, and 10c results in the total net convective flux rate.

$$\begin{aligned}\text{Net convective flux rate} \\ &= \left[\frac{\partial}{\partial x}(u C) + \frac{\partial}{\partial y}(v C) + \frac{\partial}{\partial z}(w C) \right] dx dy dz\end{aligned}\quad (11)$$

Diffusive flux rates. For net diffusive flux rate in the x -direction, Eq. 9a becomes:

$$\begin{aligned}\text{Diffusive flux rate(out - in)}_x \\ &= \frac{\partial}{\partial x} \left(-D \frac{\partial C}{\partial x} A_x \right) dx \\ &= \frac{\partial}{\partial x} \left(-D \frac{\partial C}{\partial x} \right) dx dy dz\end{aligned}\quad (12a)$$

The y - and z -directions give a result similar to Eq. 12a:

$$\begin{aligned}\text{Diffusive flux rate(out - in)}_y \\ &= \frac{\partial}{\partial y} \left(-D \frac{\partial C}{\partial y} A_y \right) dy \\ &= \frac{\partial}{\partial y} \left(-D \frac{\partial C}{\partial y} \right) dx dy dz\end{aligned}\quad (12b)$$

$$\begin{aligned}\text{Diffusive flux rate(out - in)}_z \\ &= \frac{\partial}{\partial z} \left(-D \frac{\partial C}{\partial z} A_z \right) dz \\ &= \frac{\partial}{\partial z} \left(-D \frac{\partial C}{\partial z} \right) dx dy dz\end{aligned}\quad (12c)$$

Finally, Eqs. 12a, 12b, and 12c can be added to write an equation describing the net diffusive flux rate (out-in) out of the control volume:

$$\begin{aligned}\text{Net diffusive flux rate} \\ &= - \left[\frac{\partial}{\partial x} \left(D \frac{\partial C}{\partial x} \right) + \frac{\partial}{\partial y} \left(D \frac{\partial C}{\partial y} \right) + \frac{\partial}{\partial z} \left(D \frac{\partial C}{\partial z} \right) \right] \\ &\quad dx dy dz\end{aligned}\quad (13)$$

The diffusion coefficient is often not a function of distance, such that Eq. 13 can be further simplified by putting the constant value diffusion coefficient in front of the partial derivative. However, we will also be substituting turbulent diffusion and dispersion coefficients for D when appropriate to certain applications, and they are not always constant in all directions. We will therefore leave the diffusion coefficient inside the brackets for now.

Control volume mass balance. Now Eqs. 1, 5, 6, 11, and 13 can be combined into a mass balance on our box for Cartesian coordinates. After dividing by $\mathcal{V} = dx dy dz$ and moving the diffusive flux terms to the right-hand side, this mass balance is:

$$\begin{aligned}\frac{\partial C}{\partial t} + \frac{\partial}{\partial x}(u C) + \frac{\partial}{\partial y}(v C) + \frac{\partial}{\partial z}(w C) \\ &= \left[\frac{\partial}{\partial x} \left(D \frac{\partial C}{\partial x} \right) + \frac{\partial}{\partial y} \left(D \frac{\partial C}{\partial y} \right) + \frac{\partial}{\partial z} \left(D \frac{\partial C}{\partial z} \right) \right] + S\end{aligned}\quad (14)$$

When working with a computational transport code, there is little reason to further simplify Eq. 14. One primary objective of this section, however, is to

develop approximate analytical solutions to environmental transport problems, and we will normally be assuming that diffusivity is not a function of position, or x , y , and z . The convective transport terms can be expanded with the chain rule of partial differentiation:

$$\frac{\partial}{\partial x}(uC) = u \frac{\partial C}{\partial x} + C \frac{\partial u}{\partial x} \quad (15a)$$

$$\frac{\partial}{\partial y}(vC) = v \frac{\partial C}{\partial y} + C \frac{\partial v}{\partial y} \quad (15b)$$

$$\frac{\partial}{\partial z}(wC) = w \frac{\partial C}{\partial z} + C \frac{\partial w}{\partial z} \quad (15c)$$

This may not seem like much help, because we have expanded three terms into six. However, if the flow is assumed to be incompressible, a derivation given in fluid mechanics texts (the continuity equation) is:

$$\rho \left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} \right) = 0 \quad (16)$$

where ρ is the density of the fluid. Since Eqs. 15a, b, and c are added together in the mass balance equation, the incompressible assumption means that the terms on the far right-hand side of these equations will sum to zero, or:

$$\frac{\partial}{\partial x}(uC) + \frac{\partial}{\partial y}(vC) + \frac{\partial}{\partial z}(wC) = u \frac{\partial C}{\partial x} + v \frac{\partial C}{\partial y} + w \frac{\partial C}{\partial z} \quad (17)$$

The incompressible flow assumption is most always accurate for water in environmental applications, and is often a good assumption for air. Air flow is close to incompressible as long as the Mach number (flow velocity/speed of sound) is below 0.3. A Mach number of 0.3 corresponds to an air flow velocity of approximately 110 m/s.

Equation 14 then becomes

$$\begin{aligned} \frac{\partial C}{\partial t} + u \frac{\partial C}{\partial x} + v \frac{\partial C}{\partial y} + w \frac{\partial C}{\partial z} \\ = D \left(\frac{\partial^2 C}{\partial x^2} + \frac{\partial^2 C}{\partial y^2} + \frac{\partial^2 C}{\partial z^2} \right) + S \end{aligned} \quad (18)$$

The only assumptions made in developing Eq. 18 are (1) that diffusivity does not change with spatial coordinate and (2) incompressible flow. Equation 18 will be further simplified in order to develop analytical solutions for mass transport problems. In some cases,

all that needs to be done is orient the flow direction so that it corresponds with one of the coordinate axes. There would then be only one convection term.

Adsorption and Desorption in Sediments and Soils

Sorption relates to a compound sticking to the surface of a particle. Adsorption relates to the process of compound attachment to a particle surface, and desorption relates to the process of detachment. Sorption processes will now be reviewed because there are many compounds that are sorptive and subject to spills. Then the solutions of the diffusion equation can be examined as they apply to highly sorptive compounds.

Environmental chemicals are generally classified by the Greek terms hydrophilic (likes water) and hydrophobic (hates water). Water is a polar molecule, in that it has two hydrogen atoms on one side, and an oxygen atom on the other. Solutes with a polarity or charge, therefore, will have water molecules surrounding them with the tendency to have the proper charge of atom adjacent to the solute. Most amides and alcohols are strongly polar, and also soluble in water. These are generally hydrophilic compounds. Other organic compounds with larger molecular weights, especially with aromatic rings, are generally nonpolar and are classified as hydrophobic compounds. It makes sense that these hydrophobic compounds would adsorb to the nonpolar organic material in the sediments or soils. There are handbooks [7] that can be used to estimate the chemical thermodynamics of a water-particle system.

How can sorption be handled in our transport equation? For particles that are not transported with the flow field, like sediments and groundwater flow, we are interested in the water concentrations. The sorbed portion of the compound is not in the solute phase, and should not be considered in the transport equation except when transfer of the compound between the water and particles occur. Adsorption would then be a sink of the compound and desorption would be a source.

Let us assign S_p to be the mass of chemical sorbed to particles per mass of solids contained in our control volume, and C to be the concentration of the compound in solution. Then, the source term in the

diffusion equation is equal to the rate of change of mass due to adsorption and desorption per unit volume, or:

$$S = \frac{\rho_b}{\varepsilon} \frac{\partial S_p}{\partial t} \quad (19)$$

where ρ_b is the bulk density of the solid (mass of solid/volume of fluid and solid), ε is the porosity of the media (volume of fluid/volume of fluid and solid), and $\partial S_p/\partial t$ is the rate of sorption relative to the mass of solid (mass adsorbed/mass of solid/time). If the sorption rate is negative, desorption is occurring. The units of S in Eq. 19 are mass adsorbed/volume of fluid/time. This is similar to the units for the $\partial C/\partial t$ term, which are a change of mass/volume of fluid/time.

The source term in Eq. 19 requires a separate differential equation for S_p , which would incorporate the concentration of the compound in solution. There would thus be two equations that need to be solved simultaneously. However, most sorption rates are high, relative to the transport rates in sediments and soil. Thus, *local equilibrium in adsorption and desorption is often a good assumption*. It also simplifies the solution to a transport problem considerably. If that assumption is made, S_p changes in proportion to C alone, or:

$$S_p = S_p(C) \quad (20)$$

and

$$\frac{\partial S_p}{\partial t} = \frac{\partial S_p}{\partial C} \frac{\partial C}{\partial t} \quad (21)$$

Now, if Eq. 21 is substituted into 19, we get:

$$S = \frac{\rho_b}{\varepsilon} \frac{\partial S_p}{\partial C} \frac{\partial C}{\partial t} \quad (22)$$

The $\partial S_p/\partial C$ term can be found from the equilibrium relationship of Freundlich isotherms, expressed as:

$$S_p = K_d C^\beta \quad (23)$$

where K_d is an equilibrium-partitioning coefficient between the fluid and sorption to the solid and β is a coefficient fit to measured data. Then,

$$\frac{\partial S_p}{\partial C} = \beta K_d C^{\beta-1} \quad (24)$$

At the lower concentrations normally found in the environment, $\beta = 1$ is a valid assumption. Then Eq. 24 becomes

$$\frac{\partial S_p}{\partial C} = K_d (\beta = 1) \quad (25)$$

Substituting Eq. 25 into 22 now results in a source term that no longer contains the variable S_p , and keeps the partial differential equation (PDE) of our mass balance linear:

$$S = \frac{\rho_b}{\varepsilon} K_d \frac{\partial C}{\partial t} \quad (26)$$

Now, if Eq. 26 is substituted into our mass transport Eqs. 15a–c for the source term, the result is a PDE where the only dependent variable is C :

$$\begin{aligned} \frac{\partial C}{\partial t} + u \frac{\partial C}{\partial x} + v \frac{\partial C}{\partial y} + w \frac{\partial C}{\partial z} \\ = D \left(\frac{\partial^2 C}{\partial x^2} + \frac{\partial^2 C}{\partial y^2} + \frac{\partial^2 C}{\partial z^2} \right) - \frac{\rho_b}{\varepsilon} K_d \frac{\partial C}{\partial t} \end{aligned} \quad (27)$$

or

$$\begin{aligned} \left(1 + \frac{\rho_b}{\varepsilon} K_d \right) \frac{\partial C}{\partial t} + u \frac{\partial C}{\partial x} + v \frac{\partial C}{\partial y} + w \frac{\partial C}{\partial z} \\ = D \left(\frac{\partial^2 C}{\partial x^2} + \frac{\partial^2 C}{\partial y^2} + \frac{\partial^2 C}{\partial z^2} \right) \end{aligned} \quad (28)$$

If we divide Eq. 28 by the term $(1 + K_d \rho_b/\varepsilon)$, we can see that all convective and diffusive transport is *retarded* by equilibrium adsorption and desorption. Thus, a *retardation factor* is defined:

$$R = \text{retardation factor} = 1 + K_d \rho_b/\varepsilon \quad (29)$$

and Eq. 28 becomes:

$$\begin{aligned} \frac{\partial C}{\partial t} + \frac{u}{R} \frac{\partial C}{\partial x} + \frac{v}{R} \frac{\partial C}{\partial y} + \frac{w}{R} \frac{\partial C}{\partial z} \\ = \frac{D}{R} \left(\frac{\partial^2 C}{\partial x^2} + \frac{\partial^2 C}{\partial y^2} + \frac{\partial^2 C}{\partial z^2} \right) \end{aligned} \quad (30)$$

Equation 30 indicates that as long as it can be assumed that the sorption rates are fast compared to our transport rates and the equilibrium partitioning is linearly related to concentration, the retardation factor can utilize and simply convert all of the transport terms through dividing by R . Thus, if there is a spill into the groundwater table that is highly hydrophobic, it would

transport through the soil more slowly than one which is hydrophilic. Both the convective and the diffusive flux would be “retarded” for the hydrophobic compound. If both hydrophilic and hydrophobic compounds are contained in the spill, the hydrophilic compound would show up first at a downstream location. The similarity to the manner in which a chromatographic column separates compounds is not fortuitous, because the column is separating compounds through their sorption to the column’s media.

Determination of K_d from octanol-water partitioning coefficient. There have been a number of empirical equations developed to determine the water-solid partitioning coefficient, K_d [7]. These are primarily for the many organic chemicals that exist in the environment, usually due to human impacts. Many of them use the octanol-water partitioning coefficient for the compound as an indicator of hydrophobicity. Octanol is a relatively insoluble organic compound. Since most organic compounds tend to adsorb to the organic portion of the particles, a hydrophobic organic compound placed in an octanol-water solution will tend toward the octanol. The ratio of concentration in the octanol over concentration in the water will indicate the degree of the hydrophobicity. It is a straightforward and relatively easy measurement to make, so most organic compounds of interest in the environment have an octanol-water partitioning coefficient that has been measured.

Karickhoff et al. [2] developed a simple empirical equation for equilibrium partitioning of organic compounds that will be used in this text (other equations are given in Lehman et al. [7]):

$$K_d = \beta f K_{ow} \quad (31)$$

where K_{ow} is the dimensionless octanol-water partitioning coefficient, f is the fraction of soil that is organic matter (usually from zero in sand to 0.01 in sandy soil to 0.10 in muck), and β is an empirical coefficient, estimated by Karickhoff to be $0.41 \text{ cm}^3/\text{g}$. It is generally the organic matter in the medium to which organic compounds adsorb, hence the use of organic fraction.

The other parameters required to compute a retardation coefficient are the bulk density, ρ_b , and the porosity of the media, ε . The bulk density of the water and soil is typically $1.6\text{--}2.1 \text{ g/cm}^3$. The porosity

of the soil or sediments is typically $0.2\text{--}0.4$. Thus ρ_b/ε is typically between 4 and 10 g/cm^3 .

Example Applications of the Diffusion Equation

The first application of the diffusive is transport of oxygen into lake sediments and the use of oxygen by the bacteria to result in a steady-state oxygen concentration profile.

Example 1: Steady O_2 concentration profile in lake sediments (steady-state solution with a first-order sink). Given a concentration, C_0 , in the overlying water, and a first-order sink of oxygen in the sediments, develop an equation to describe the dissolved oxygen concentration profile in the sediments (Fig. 5).

Assume:

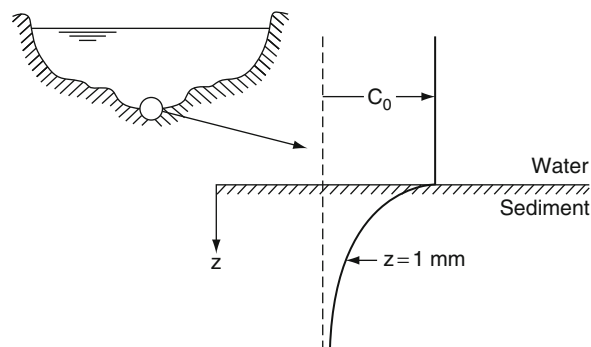
- Steady: $\frac{\partial}{\partial t} \Rightarrow 0$
- No flow: $u, v, w \Rightarrow 0$
- Small horizontal variation: $\frac{\partial^2 C}{\partial z^2} \gg \frac{\partial^2 C}{\partial x^2}, \frac{\partial^2 C}{\partial y^2}$
- No sorption: $R = 1$ (accurate for O_2 in sediments)
- First-order sink: $S = -kC$, where k is a rate constant

Then, the diffusive mass transport Eqs. 15a–c becomes:

$$0 = D \frac{\partial^2 C}{\partial z^2} - kC$$

or, since $C = C(z)$

$$0 = D \frac{d^2 C}{dz^2} - kC$$



Transport in the Environment. Figure 5

Illustration of dissolved oxygen profile in lake sediments. (From [1])

A solution to this equation requires two boundary conditions because it is a second-order equation. These two are:

$$\text{B.C. \#1: } @ z = 0, C = C_0$$

$$\text{B.C. \#2: } @ z \rightarrow \infty, C \Rightarrow 0$$

This solution may be achieved by: (1) separating variables and integrating or (2) solving the equation as a second-order, linear ordinary differential equation (ODE). The latter will be used since the solution technique is more general.

1. Assign λ to be the $\frac{d}{dz}$ operator. Then, the equation becomes

$$\left(\lambda^2 - \frac{k}{D} \right) C = 0$$

2. Solve for λ

$$\lambda = \pm \sqrt{k/D}$$

3. The solution, developed in texts on solving ordinary differential equations [6], is

$$C = \beta_1 e^{\lambda_1 z} + \beta_2 e^{\lambda_2 z} \quad \lambda_1 = +\sqrt{k/D}$$

$$\lambda_2 = -\sqrt{k/D}$$

4. β_1 and β_2 are determined from boundary conditions

Apply B.C. #2:

$$C = 0 = \beta_1 e^{\sqrt{k/D}\infty} + \beta_2 e^{-\sqrt{k/D}\infty}$$

This is only possible if $\beta_1 = 0$. Apply B.C. #1:

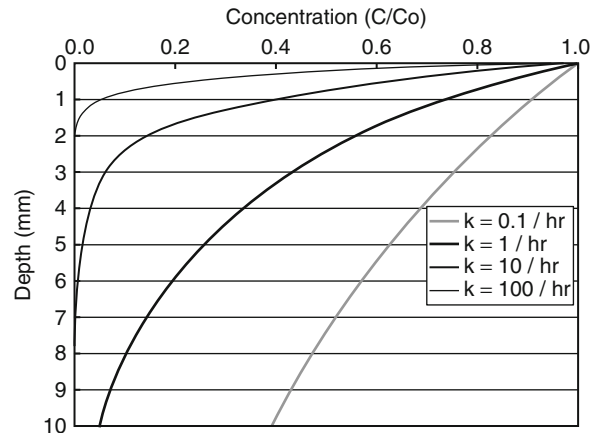
$$C_0 = 0 + \beta_2 e^{-0} = \beta_2$$

Thus, the solution is:

$$C = C_0 e^{-\sqrt{k/D}z}$$

which is plotted in Fig. 6.

At steady state, the oxygen profile is a balance between diffusion from the sediment surface and bacterial use of oxygen in the sediments. If the sediments are mostly sand, the depth of the layer with oxygen can be 10 cm or more. If the sediments have a substantial organic content (like a mud), the aerobic layer ($>0.1 \text{ g/m}^3$ oxygen concentration) can be less than 1 mm.



Transport in the Environment. Figure 6
Solution to Example 1

Example 2: Unsteady dissolution of a highly soluble pollutant (Herbicides, Pesticides, Ammonia, Alcohols, etc.) into groundwater (unsteady, one-dimensional solution with pulse boundary conditions). A tanker truck carrying a highly soluble compound in Mississippi tried to avoid an armadillo at night, ran off the interstate at a high speed, turned over in the drainage ditch, and spilled a soluble compound. The compound has infiltrated into the ground, and much of it has reached and temporarily spread out over the groundwater table, as illustrated in Fig 7. As part of a spill response team, you need to estimate the groundwater contamination. Predict concentrations over time in the groundwater table.

The mass transport equation for this example is:

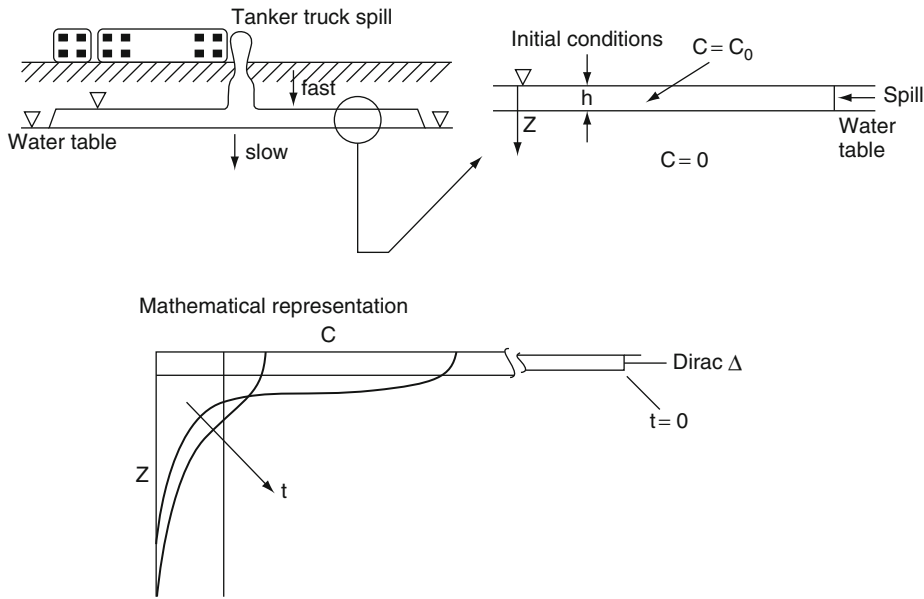
$$\begin{aligned} \frac{\partial C}{\partial t} + u \frac{\partial C}{\partial x} + v \frac{\partial C}{\partial y} + w \frac{\partial C}{\partial z} \\ = D \left(\frac{\partial^2 C}{\partial x^2} + \frac{\partial^2 C}{\partial y^2} + \frac{\partial^2 C}{\partial z^2} \right) + S \end{aligned}$$

Assume:

1. Minimal horizontal variations

$$0 \cong \frac{\partial C}{\partial x} = \frac{\partial^2 C}{\partial x^2} \cong \frac{\partial C}{\partial y} = \frac{\partial^2 C}{\partial y^2}$$

2. No flow in the vertical direction, $w = 0$
3. No reactions, including adsorption and desorption, such that $S = 0$.



Transport in the Environment. Figure 7
Illustration of the tanker truck spill. (From [1])

Then with these three assumptions, the governing equation becomes:

$$\frac{\partial C}{\partial t} = D \frac{\partial^2 C}{\partial z^2}$$

The initial conditions will be simulated with these boundary conditions:

1. The mass of chemical is assumed to be spread instantaneously across a very thin layer at $t = 0$ (a Dirac delta in z and t). At $z = 0^+$, $t = 0$, the total mass = M and the total surface area is A .
2. At $z \Rightarrow \infty$, $C \Rightarrow 0$.

The above equation, with boundary conditions (1) and (2), has the solution:

$$C = \frac{2M/A}{\sqrt{4\pi Dt}} e^{-z^2/4Dt}$$

What does the solution look like? The solution can be made dimensionless by assuming that the initial thickness of the spill layer is Δh . Then, a new variable $z = k \Delta h$ will be used in assigning:

$$\eta = \frac{\Delta h}{\sqrt{4Dt}}$$

with

$$C^* = \frac{CA \Delta h}{2M}$$

Substituting these equations into the solution gives:

$$C^* = \frac{\eta}{\sqrt{\pi}} e^{-(k\eta)^2}$$

which is plotted versus depth at various times in Fig. 8. The concentration at $z = 0$ decreases as the initial mass is diffused. At low values of time, the concentration at and close to $z = 0$ is strongly dependent upon the Δh chosen. At larger times and deeper depths, however, this dependency decreases, and the solution becomes independent of Δh .

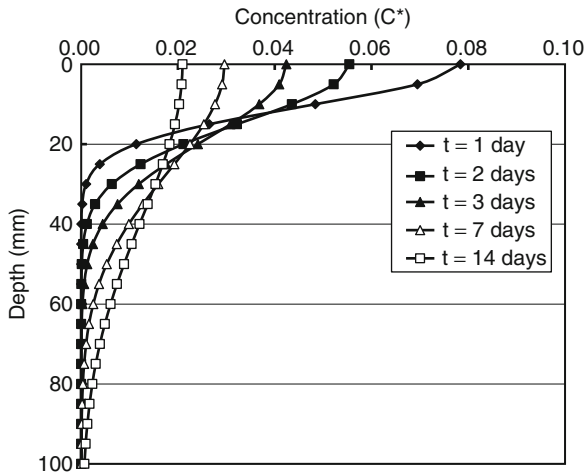
It is interesting to note that the solution is very similar to a Gaussian probability distribution, with the following relationship for $P(z)$:

$$P(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(z-z_m)^2/2\sigma^2}$$

where z_m is the depth of the maximum concentration (or the center of concentration mass).

Comparing the probability distribution and the solution to this problem, we can see that:

$$2\sigma^2 \Leftrightarrow 4Dt$$



Transport in the Environment. Figure 8

Solution to the tanker truck spill illustrating groundwater concentration versus distance at various times. $\Delta h = 2 \text{ mm}$, $D = 6 \times 10^{-4} \text{ mm}^2/\text{s}$

or:

$$D = \sigma^2 / 2t$$

Note that if we measure σ , we can determine D .

Example 3: Dichlorobenzene concentration in lake sediments due to a plating facility discharge (solution to a concentration front). Sometimes the boundary conditions can be approximated as a step in concentration. This difference in boundary conditions changes the solution from one which is related to pulse boundaries (known mass release) to one resulting from a concentration front with a known concentration at one boundary.

For many years, a plating facility for a telecommunications company let their rinse waters flow into an adjacent lake. The compounds used in their rinse included dichlorobenzene, which is a semi-volatile compound that also has a fairly high tendency to adsorb to organic compounds in the sediments. Within a few years of the plating facility opening, the dichlorobenzene concentration reached a steady-state value in the lake waters as illustrated in Fig. 9. Estimate the buildup of dichlorobenzene in the sediments during the 50 years since the facility opened until it stopped discharging its untreated waste water.

Assumptions:

1. Biodegradation is small. $\therefore S \Rightarrow 0$ except for sorption.
2. Variation in x and y are small

$$\frac{\partial^2 C}{\partial x^2}, \frac{\partial^2 C}{\partial y^2} \ll \frac{\partial^2 C}{\partial z^2}$$

3. No flow in sediments under lake: $u = v = w = 0$
4. $D \cong 6 \times 10^{-10} \text{ m}^2/\text{s}$
5. $\frac{\rho_b}{\varepsilon} = 6.3 \Rightarrow R = 1 + \frac{\rho_b}{\varepsilon} K_d = 96$

Then the diffusion equation for the sediments becomes:

$$\frac{\partial C}{\partial t} = \frac{D}{R} \frac{\partial^2 C}{\partial z^2}$$

with boundary conditions:

- (a) $t > 0, z = 0; C = C_0$
- (b) $t = 0, z \neq 0; C = 0$

There are three known techniques to solve this governing equation:

Laplace transforms, Fourier transforms, and change of variables, which incorporates both luck and skill. We will use change of variables:

$$\text{Assign } \eta = \frac{z}{\sqrt{4Dt/R}}$$

$$\frac{\partial C}{\partial t} = \frac{\partial C}{\partial \eta} \frac{\partial \eta}{\partial t} = \frac{-1}{4} \frac{z}{\sqrt{Dt/R}} \frac{\partial C}{\partial \eta} = \frac{-\eta}{2t} \frac{\partial C}{\partial \eta}$$

$$\frac{\partial C}{\partial z} = \frac{\partial C}{\partial \eta} \frac{\partial \eta}{\partial z} = \frac{1}{2\sqrt{Dt/R}} \frac{\partial C}{\partial \eta}$$

$$\frac{\partial^2 C}{\partial z^2} = \frac{\partial}{\partial \eta} \left(\frac{\partial C}{\partial z} \right) \frac{\partial \eta}{\partial z} = \frac{R}{4Dt} \frac{\partial^2 C}{\partial \eta^2}$$

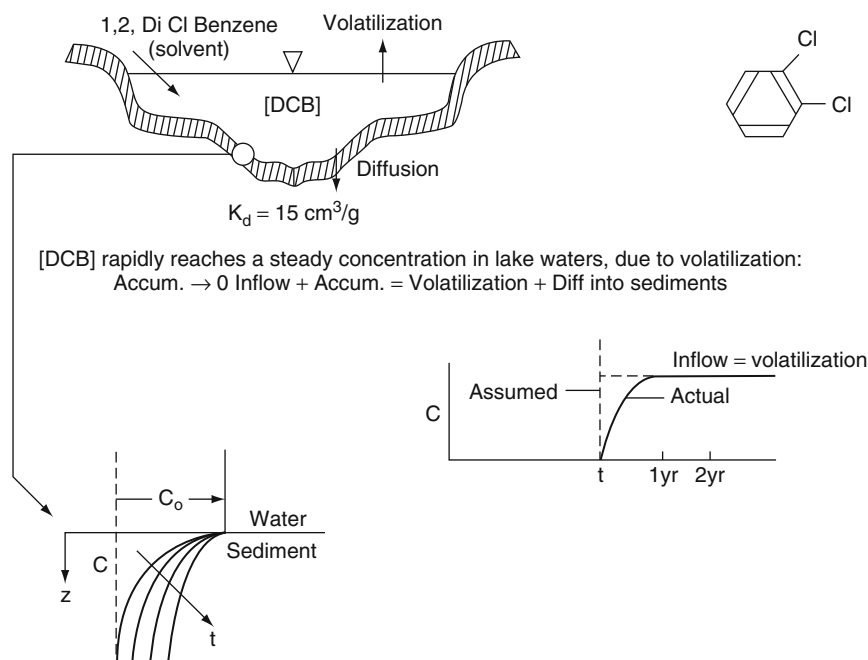
Then the governing equation becomes:

$$\frac{-\eta}{2t} \frac{\partial C}{\partial \eta} + \frac{D}{R} \left(\frac{R}{4Dt} \frac{\partial^2 C}{\partial \eta^2} \right) = 0$$

or

$$\frac{d^2 C}{d\eta^2} + 2\eta \frac{dC}{d\eta} = 0 \quad (32)$$

This equation may be written as: $\frac{dC'}{d\eta} + 2\eta C' = 0$



Transport in the Environment. Figure 9

Illustration of concentration front moving down into the sediments of a lake. (From [1])

where $C' = \frac{dC}{d\eta}$ or

$$\frac{1}{C'} dC' = -2\eta d\eta$$

We can integrate this:

$$\ln C' = -\eta^2 + \beta_0$$

or

$$C' = e^{\beta_0} e^{-\eta^2} = \beta_1 e^{-\eta^2}$$

Now integrate again:

$$C = \beta_1 \int_0^\eta e^{-\eta^2} d\eta + \beta_2 \Rightarrow \beta_1 \int_0^\eta e^{-\phi^2} d\phi + \beta_2$$

Now, note that the error function is given as:

$$\text{erf}(\eta) = \frac{2}{\sqrt{\pi}} \int_0^\eta e^{-\phi^2} d\phi$$

and the complementary error function is $\text{erfc}(\eta) = 1 - \text{erf}(\eta)$. Values of the error function and complementary error function for various values of η may be found in an Internet search. The error function is designed

such that $\text{erf}(\infty) = 1$, $\text{erfc}(\infty) = 0$, $\text{erf}(0) = 0$, and $\text{erfc}(0) = 1$. The solution may therefore be written as:

$$C = \beta_1 \text{erf}(\eta) + \beta_2$$

Now we need to determine our boundary conditions in terms of η :

1. $t > 0, z = 0, \eta = 0, C = C_0$
2. $t = 0, z = 0, \eta = \infty, C = 0$

Checking other boundary conditions:

$$t \rightarrow \infty, \eta \Rightarrow 0, C = C_0$$

$$z \rightarrow \infty, \eta \Rightarrow \infty, C = 0$$

Now, at $\eta = 0, C = C_0$, thus:

$$C_0 = \beta_1 0 + \beta_2$$

or

$$\beta_2 = C_0$$

At $\eta = \infty, C = 0$

$$0 = \beta_1 1 + C_0$$

or

$$\beta_1 = -C_0$$

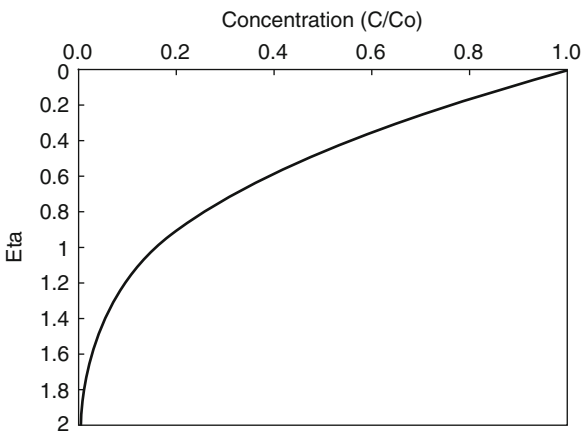
Then, our solution is:

$$C = C_0 \left(1 - \operatorname{erf} \left(\frac{z}{\sqrt{4Dt/R}} \right) \right) = C_0 \operatorname{erfc}(\eta) \quad (33)$$

which is illustrated in Fig. 10.

We will now apply Eq. 33 to estimate the dichlorobenzene penetration versus time from spillage. The results are given in Table 2, which gives the interstitial dichlorobenzene concentrations.

The total concentration (TDCB(z)) includes compound adsorbed to the sediments.



Transport in the Environment. Figure 10
Illustration of the effect of the change of variables used in Example 3. $\text{Erf} = \eta$

Transport in the Environment. Table 2 Penetration of dichlorobenzene into the sediment over time

Time, z	1 year, C/C_0	4 years, C/C_0	10 years, C/C_0	50 years, C/C_0
1 mm	0.96	0.98	0.988	0.994
1 cm	0.62	0.803	0.87	0.94
10 cm	0	0.015	0.11	0.48
20 cm	0	0	0.01	0.16
30 cm	0	0	0	0.03
100 cm	0	0	0	0

$$\text{TDCB}(z) = C(z) + (1 - \varepsilon)\rho_s S$$

where ε = porosity

$1 - \varepsilon$ = % by volume sediment $\cong 0.6$

ρ_s = density of sediment $\cong 2.5 \text{ g/cm}^3$

S = concentration of sorbed compound (g DCB/g sediment)

Since $S = K_d C_1$, the above equation becomes:

$$\begin{aligned} \text{TDCB}(z) &= C(z)(1 + (1 - \varepsilon)\rho_s K_d) \\ &= C(z)(1 + 0.6(2.5 \text{ g/cm}^3)(15 \text{ cm}^3/\text{g})) \end{aligned}$$

or

$$\text{TDCB} = 23.5 C(z)$$

Thus, the total dichlorobenzene per volume of sediment and water would be 23.5 times the concentrations given in Table 2.

Conclusion

The purpose of this section of the volume is to introduce the reader to the equations and mathematics used in developing approximate solutions (due to simplified boundary conditions) to convection-diffusion processes. One may say that, with computational capabilities, there is no longer any need to develop these approximate solutions. However, these approximate solutions are useful in the following manners:

1. A quick, back of the envelope solution is always much quicker and more reliable than a computational solution. Computational solutions require substantial time to develop and are often wrong until they are fully vetted.
2. A computational solution always requires vetting, which means that a computational solution is compared to an analytical solution, hopefully in a similar condition with simplified boundary conditions. This means that some analytical solution is always needed, and as close to the real simulation as possible, to make sure that the computational solution is doing what the user desires.
3. Developing analytical solutions are an excellent means of getting a feel for solutions to the convection-diffusion equations. It is a knowledge-building practice that is difficult to surpass.

Note that the examples given in this section do not include any with convection. That is because convection in the environment most always includes either turbulence (surface waters and the atmosphere) or dispersion (groundwater). These will be dealt with in other sections.

Future Directions

Most of the future directions with regard to transport in the environment (without turbulent transport) will involve transport across interfaces, such as the air–water and solid–fluid interfaces. While research has been conducted on describing the predominant transport mechanisms for these two cases (McCready et al. [8], [3–5]), there is more to be done. An especially vexing problem is transport in the vadose zone of soils (unsaturated zone). The multiplicity of three interfaces, air, water, and soil, and the heterogeneities in the soil make this a complex problem to handle in a deterministic manner. However, meaningful relationships for an effective diffusion coefficient still need to be developed.

Bibliography

Primary Literature

1. Gulliver JS (2007) An introduction to chemical transport in the environment. Cambridge University Press, Cambridge
2. Karickhoff SW, Brown DS, Scott TA (1979) Sorption of hydrophobic pollutants on natural sediments. *Water Res* 13:241
3. Brunley BH, Jirka GH (1987) Near-surface turbulence in a grid-stirred tank. *J Fluid Mech* 183:235–263
4. Campbell JA, Hanratty TJ (1982) Mass transfer between a turbulent fluid and a solid boundary: linear theory. *AIChE J* 28:988
5. Janzen J, Jirka H, Jirka G, Schulz H, Gulliver JS (2010) Estimation of mass transfer velocity based on measured turbulence parameters. *Am Inst Chem Eng J* 56(8):2005–2017
6. Kreyszig E (1982) Advanced engineering mathematics, 4th edn. Wiley, New York
7. Lehman WJ, Reehl WF, Rosenblatt DH (1990) Handbook of chemical property estimation. American Chemical Society, Washington, DC
8. McCready MA, Vassiliadou E, Hanratty T (1986) Computer simulation of turbulent mass transfer at a mobile interface. *AIChE J* 32(7):1108
9. Fick AE (1855) Ueber Diffusion, *Annalen der Physik* 170(1):59–86

Books and Reviews

- Crank J (1975) The mathematics of diffusion, 2nd edn. Oxford University Press, Oxford
- Cussler EL (1997) Diffusion: mass transfer in fluid systems, 2nd edn. Cambridge University Press, Cambridge

Transport with Jets and Plumes of Chemicals in the Environment

WENMING ZHANG, NALLAMUTHU RAJARATNAM,
DAVID Z. ZHU

Department of Civil and Environmental Engineering,
University of Alberta, Edmonton, AB, Canada

Article Outline

Glossary
Definition of the Subject and Its Importance
Introduction
Turbulent Jets and Plumes in Stagnant Environment
Effect of Boundaries on Jets and Plumes
Multiphase Jets and Plumes
Future Directions
Bibliography

Glossary

Bubbly jet The jet produced by injecting gas-liquid mixture into a liquid.

Buoyant jet The plume with momentum or jet with buoyancy.

Circular jet The jet produced through a nozzle with a circular cross section.

Diffuser The device which has multiple nozzles to quickly mix the discharged substances (e.g., effluent or air) with the surrounding ambient fluid.

Jet The flow generated by the release of momentum usually through a nozzle or slot.

Jet in coflow The jet discharged in the direction of a flowing ambient fluid.

Jet in crossflow The jet discharged at an oblique angle to a flowing ambient fluid.

Plane jet Also called “slot jet” or “two-dimensional jet”, the jet produced through a slender slot.

Plume The flow generated by the release of buoyancy.

Surface jet The jet discharged at (or near) the surface of an ambient fluid.

Slurry jet The jet produced by injecting solid-liquid mixture into a liquid.

Wall jet The jet discharged tangentially or at a certain angle to a solid boundary (wall).

Definition of the Subject and Its Importance

Jets and plumes are common in our environment. Some examples of jets are: wastewater discharged from an outfall, emission from an aircraft or vehicle, and the eruption of volcano. Some examples of plumes are: the smoke from a chimney stack or cigarette, the thermal plumes from a fire, municipal wastewater or hot water discharged in deep water, and oil spill from sea bed. One of the most important features of jets or plumes is its ability of entraining ambient fluid to achieve self-dilution. This greatly triggers our interests to study jets and plumes. This book chapter is a review of the studies on turbulent jets and plumes, with a focus on the transport of conservative pollutants.

Introduction

The earliest experimental study of turbulent jets appears to be the work of Trupel on circular jets in 1915 [1]. Förthmann performed an experimental study of plane turbulent jets in 1934 and his work also considered plane turbulent wall jets [23]. The results of these investigations showed the similarity of the velocity profiles at different distances from the sources of the jets. These studies were followed by the extensive investigations of Hinze and Zijnen [28] on circular jets and Albertson et al. [4] for plane and circular jets. Turbulence characteristics of plane jets were studied by Heskestad [27] for plane jets and Wagnanski and Fielder [105] for circular jets. Theoretical solutions for plane jets were developed by Tollmien in 1926 for plane jets and by Goertler in 1942 for circular jets [73]. Numerical studies of turbulent jets followed, starting with the work of Rodi and Spalding [82]. Abramovich [1], Rajaratnam [73], and Fischer et al. [21] provided comprehensive treatment of jets.

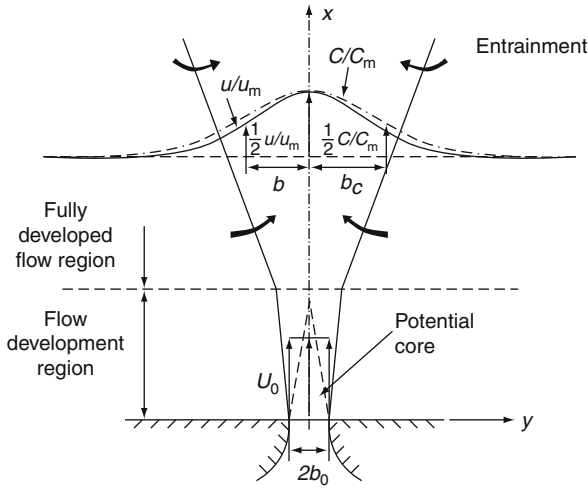
Turner [97] provides an introduction to study turbulent plumes in his book on *Buoyancy Effects in Fluids*. Rouse et al. [84] performed an experimental study of plane and circular plumes wherein they found that the

velocity and density defect profiles were similar if proper scales were chosen for velocity, width, and density defect. Morton et al. [58] published an integral study of plumes, wherein the concept of entrainment coefficient was introduced. Since then numerous studies have been conducted on turbulent plumes and forced turbulent plumes (buoyant jets). These studies have been summarized in Chen and Rodi [15] and Lee and Chu [45]. Turbulent jets and plumes have been studied extensively not only because these flows are very interesting but also that they are of considerable practical importance in the fields of hydraulic, mechanical, aeronautical, environmental, and chemical engineering and many other fields.

This book chapter will first review the most classic and well-established theories on simple jets or plumes in stagnant water, and then consider effects of boundaries including: the bed (wall) and the surface of ambient fluid; coflowing and cross-flowing ambient fluid; and the interaction of neighboring jets in the case of multiple jets. Next, two kind of multiphase jets and plumes – bubbly jets and plumes, and slurry jets and plumes – will be briefly introduced. Multiphase jets and plumes are much more complicated compared to the single-phase ones, but have gained more interest in recent years because of their wide applications. Finally, some directions for future researches will be highlighted.

Turbulent Jets and Plumes in Stagnant Environment

In this section, the focus is on the transport of conservative pollutants in a steady-state turbulent jet or plume issuing from a simple (plane or circular) nozzle into stagnant ambient fluid of large extent. Such jets or plumes are called simple jets or simple plumes. Theories in this area have been well established. Close to the nozzle exit, there is a wedge-like or cone-like region termed “potential core” where the width of initial velocity distribution decays to a point. The length of the potential core is very short, about $10.4b_0$ for a plane jet where b_0 is the half slot width, or $6.2d_0$ for a circular jet where d_0 is the nozzle diameter [45]. Therefore, for practical purposes, our attention will be limited on the flow beyond the potential core, i.e., in the “fully developed flow” region (see Fig. 1).



Transport with Jets and Plumes of Chemicals in the Environment. Figure 1

Schematic of a simple plane jet (for circular jet, replace $2b_0$ by d_0 and y by r)

The reader who is interested in flow development region may refer to Rajaratnam [73].

Simple Jets

The integral method is the most common method for analyzing simple jets or plumes. The following is a brief introduction on this method. For a plane jet as shown in Fig. 1, the Reynolds-averaged Navier-Stokes equation in x direction, continuity equation, and pollutant conservation equation, respectively, can be simplified as [73]:

$$u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} = \frac{1}{\rho} \frac{\partial \tau}{\partial y} \quad (1)$$

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0 \quad (2)$$

$$\frac{\partial uC}{\partial x} + \frac{\partial vC}{\partial y} = \varepsilon \frac{\partial^2 C}{\partial y^2} \quad (3)$$

where u and v are the time-averaged velocities in x and y directions, respectively; C is the pollutant concentration; ρ is the density of the fluid; τ is the turbulent shear stress and ε is the mean value of turbulent diffusion coefficient. After multiplying Eq. 1 by ρ and then integrating from $y = 0$ to $y = \infty$, Eq. 1 becomes:

$$\frac{d}{dx} \int_0^\infty \rho u^2 dy = 0 \quad (4)$$

Equation 4 states that the jet momentum flux at different x -sections is conserved. Using Eq. 2 and integrating the first term of Eq. 2 from $y = 0$ to $y = \infty$, we have:

$$\frac{dq_x}{dx} = \frac{d}{dx} \int_0^\infty u dy = -v_{y=\infty} \quad (5)$$

where q_x is the jet volume flux per unit slot length. The entrainment hypothesis assumes that the entrainment velocity $v_e = v_{y=\infty} = -\alpha_e u_m$, where α_e is called the jet entrainment coefficient, u_m is jet centerline (maximum) velocity, and the negative sign indicates the ambient fluid is entrained into the jet. Equation 5 says that the jet volume flux increases with traveling distance x due to the entrainment of ambient fluid, which explains the ability of jet in diluting pollutants. Similarly, integrating Eq. 3 from $y = 0$ to $y = \infty$, we have:

$$\frac{d}{dx} \int_0^\infty uC dy = 0 \quad (6)$$

Equation 6 states that the mass fluxes of pollutants at different x -sections are conserved if the chemical or biological reactions of pollutants are not considered.

Numerous laboratory experiments and numerical simulations have confirmed that beyond the potential core, the jet velocity or concentration exhibits self-similarity. The most widely used expression for such similarity is the Gaussian distribution, which represents laboratory data satisfactorily:

$$\frac{u}{u_m} = \exp \left[-k_1 \left(\frac{y}{b} \right)^2 \right] \quad (7)$$

$$\frac{C}{C_m} = \exp \left[-k_1 \left(\frac{y}{k_2 b} \right)^2 \right] \quad (8)$$

where C_m is the jet time-averaged centerline (maximum) concentration, b is the jet velocity half-width where the velocity is 50% (if $k_1 = 0.693$, refer to Fig. 1) or 37% (if $k_1 = 1$) of u_m , $k_2 b$ defines the jet concentration half-width where the concentration is 50% or 37% of C_m , and k_2 is the ratio of concentration half-width to velocity half-width. Using Eqs. 4–8, the analytical solutions for plane jets can be derived, as shown in Table 1. The coefficients of the solutions are mainly determined from experimental results and integrations.

Transport with Jets and Plumes of Chemicals in the Environment. Table 1 Summary of mean properties of simple plane jet and circular jet

Parameter	Plane jet	Circular jet
Maximum (centerline) concentration C_m	$\frac{C_m}{C_0} = \frac{\alpha_1}{\sqrt{\frac{x}{b_0}}}$, where $\alpha_1 = 3.37$ in [21], 3.45 in [74], 3.21 in [45]	$\frac{C_m}{C_0} = \frac{\alpha_1}{\frac{x}{d_0}}$, where $\alpha_1 = 4.96$ in [21], 5.34 in [74], 5.26 in [45]
Cross-sectional average concentration C_{avg}	$\frac{C_m}{C_{avg}} = \alpha_2$, where $\alpha_2 = 1.2$ in [21], 1.25 in [45]	$\frac{C_m}{C_{avg}} = \alpha_2$, where $\alpha_2 = 1.4$ in [21], 1.76 in [74], 1.68 in [45]
Maximum (centerline) velocity u_m	$\frac{u_m}{U_0} = \frac{\alpha_3}{\sqrt{\frac{x}{b_0}}}$, where $\alpha_3 = 3.50$ in [73], 3.41 in [21], 3.65 in [45]	$\frac{u_m}{U_0} = \frac{\alpha_3}{\frac{x}{d_0}}$, where $\alpha_3 = 6.3$ in [73], 6.2 in [21] and [45], 6.13 in [74]
Velocity half-width b^a	$b = \alpha_4 x$, where $\alpha_4 = 0.10$ in [73], 0.116 in [21], 0.097 in [74], 0.12 in [45]	$b = \alpha_4 x$, where $\alpha_4 = 0.10$ in [73], 0.107 in [21], 0.096 in [74], 0.114 in [45]
Concentration half-width b_c^a	$\frac{b_c}{b} = \alpha_5$, where $\alpha_5 = 1.35$ in [21] and [45], 1.17 in [74]	$\frac{b_c}{b} = \alpha_5$, where $\alpha_5 = 1.19$ in [21], 1.17 in [74], 1.2 in [45]
Entrainment coefficient α_e	$\alpha_e = 0.053$ in [73] and [45]	$\alpha_e = 0.026$ in [73], 0.028 in [74], 0.057 in [45]

Source: [21, 45, 73, 74].

^aIn [21] and [45], b (or b_c) is defined as where the velocity (or concentration) is 37% of u_m (or C_m); while in others, defined as 50% of u_m (or C_m).

Using the same procedures as above, the equations of momentum flux, volume flux, and pollutant mass flux for a circular jet can be derived. These equations suggest that the jet momentum flux at any x -section is conserved and the jet volume flux across any x -section increases with traveling distance due to the entrainment of ambient fluid. Although the entrainment causes the pollutant to get diluted within the jet core, the mass flux of any conservative pollutant at any x -section is conserved. The solutions of these equations for circular jets are summarized in Table 1.

Simple Plumes

A plume is produced from a steady discharge of a fluid whose motion is controlled by its buoyancy, with negligible effect of initial momentum. First, a plane plume of density ρ_0 issued into a stagnant unstratified ambient fluid of density ρ_a is considered. It is assumed that $\frac{\Delta\rho}{\rho_a} \ll 1$ (true for most practical cases), where the initial density defect $\Delta\rho_0 = \rho_a - \rho_0$. After some manipulation, it can be shown that the Reynolds-averaged Navier-Stokes equation in x direction becomes:

$$u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} = \frac{1}{\rho_a} \frac{\partial \tau}{\partial y} + g \frac{\Delta\rho}{\rho_a} \quad (9)$$

where $\Delta\rho = \rho_a - \rho$ and ρ is the plume density. For a plane plume, the continuity equation and pollutant conservation equation can be simplified in the same form as for a plane jet (Eqs. 2 and 3).

Using the integral method, Eq. 9 can be reduced to:

$$\frac{d}{dx} \int_0^\infty \rho u^2 dy = \int_0^\infty g \Delta\rho dy \quad (10)$$

Equation 10 says that axial momentum flux increases in x direction, and the increase rate is equal to the buoyancy per unit length (in x direction) of the plume. For a plane plume, the continuity equation and pollutant conservation equation can be reduced the same as Eqs. 5 and 6. Equation 5 states that the volume flux of a plume increases due to the entrainment of ambient fluid, and Eq. 6 states that although the concentration of pollutant decreases, its total mass flux is conserved. Note that C in Eq. 6 can be also interpreted as $g\Delta\rho$, and then Eq. 6 becomes the integral form of buoyancy conservation equation.

Experimental results show that: similarly as for jets, the plume velocity or concentration also exhibit self-similarity and the Gaussian distribution can well describe it. Using the Gaussian profiles (Eqs. 7 and 8), the analytical solutions for a plane plume can be

derived as shown in Table 2. The coefficients differ slightly in different references as they are determined using the results of different experiments. Here, to constitute the solutions, an useful dimensionless parameter is introduced – the densimetric Froude number at the slot exit $F_0 = \frac{U_0}{\sqrt{g \left(\frac{\Delta \rho}{\rho_a} \right) b_0}}$ (for circular plume, $F_0 = \frac{U_0}{\sqrt{g \left(\frac{\Delta \rho}{\rho_a} \right) d_0}}$).

For a circular plume, the Reynolds-averaged Navier-Stokes equations, continuity equation and pollutant conservation equation, can be simplified, and the integral method can be used to obtain their integral forms of the equations. These equations indicate that the momentum flux in the plume increases with axial (x) direction, and the increase rate is equal to the buoyant force per unit axial length; the volume flux in the plume increases due to the entrainment of ambient fluid; and the pollutant mass flux (or the flux of density defect) remains invariant in the axial direction. The analytical solutions for a circular plume are also presented in Table 2. The constants in these equations

change slightly in different references where different experimental results were used.

Buoyant Jets

For a buoyant jet, the initial momentum flux cannot be neglected. Near the nozzle (slot) exit, it is expected that the buoyant flow will be like a jet; after some distance from the exit where the increase of momentum flux is much larger than the initial momentum flux, the buoyant flow will behave like a plume. For a plane buoyant jet, based on dimensional analysis, a characteristic length scale to judge whether the buoyant flow behaves like a jet or plume may be defined as:

$$L_M = \frac{M_0}{B_0^{\frac{2}{3}}} \quad (11)$$

where the initial specific momentum flux $M_0 = qU_0$; the initial specific buoyancy flux $B_0 = qg \frac{\Delta \rho_0}{\rho_a}$. If the jet centerline trajectory $l \ll L_M$, the buoyant jet can be treated as a pure jet and the simple jet equations can

Transport with Jets and Plumes of Chemicals in the Environment. Table 2 Summary of mean properties of plane plume and circular plume

Parameter	Plane plume	Circular plume
Maximum (centerline) concentration C_m or density defect $\Delta \rho_m$	$\frac{C_m}{C_0} = \frac{\Delta \rho_m}{\Delta \rho_0} = \frac{\alpha_1 F_0^{\frac{2}{3}}}{x/b_0}$, where $F_0 = \frac{U_0}{\sqrt{g \frac{\Delta \rho}{\rho_a} b_0}}$, $\alpha_1 = 3.78$ in [21], 3.84 in [74], 4.25 in [45]	$\frac{C_m}{C_0} = \frac{\Delta \rho_m}{\Delta \rho_0} = \frac{\alpha_1 F_0^{\frac{2}{3}}}{\left(\frac{x}{d_0}\right)^{\frac{2}{3}}}$, where $F_0 = \frac{U_0}{\sqrt{g \frac{\Delta \rho}{\rho_a} d_0}}$, $\alpha_1 = 7.75$ in [21], 7.83 or 9.37 in (from different methods in Ref. [74]), 8.90 in [45]
Cross-sectional average concentration C_{avg}	$\frac{C_m}{C_{avg}} = \alpha_2$, where $\alpha_2 = 1.32$ in [74], 1.25 in [45]	$\frac{C_m}{C_{avg}} = \alpha_2$, where $\alpha_2 = 1.40$ in [21], 1.70 in [45]
Maximum (centerline) velocity u_m	$\frac{u_m}{U_0} = \frac{\alpha_3}{F_0^{\frac{2}{3}}}$, where $\alpha_3 = 2.09$ in [21], 2.52 in [74], 2.85 in [45]	$\frac{u_m}{U_0} = \frac{\alpha_3}{F_0^{\frac{2}{3}} \left(\frac{x}{d_0}\right)^{\frac{1}{3}}}$, where $\alpha_3 = 4.34$ in [21], 4.00 or 4.33 in (from different methods in Ref. [74]), 4.35 in [45]
Velocity half-width b^a	$b = \alpha_4 x$, where $\alpha_4 = 0.116$ in [21] and [45], 0.128 in [74]	$b = \alpha_4 x$, where $\alpha_4 = 0.100$ in [21], 0.085 in [74], 0.105 in [45]
Concentration half-width b_c^a	$b_c/b = \alpha_5$, where $\alpha_5 = 1.35$ in [21] and [45], 1.17 in [74]	$b_c/b = \alpha_5$, where $\alpha_5 = 1.20$ in [21], 1.16 in [74], 1.19 in [45]
Entrainment coefficient α_e	$\alpha_e = 0.136$ in [74], 0.103 in [45]	$\alpha_e = 0.047$ in [74], 0.088 in [45]

Source: [21, 45, 74].

^aIn [21] and [45], b (or b_c) is defined as where the velocity (or concentration) is 37% of u_m (or C_m); while in [74], defined as 50% of u_m (or C_m).

be used; if $l \gg l_M$, it can be treated as a pure plume and the simple plume equations are valid. Buoyant jets will be plume-like beyond $\frac{l}{l_M} \geq 4 \sim 5$ [41, 65].

Characteristic length scale of jet/plume can only roughly help us calculate the evolution of a buoyant jet. The Reynolds equations, continuity equation and pollutant mass conservation equations can be simplified and the integral method can be used to obtain the analytical solutions. For a plane buoyant jet, the integral forms of the momentum, continuity, pollutant mass (or buoyancy) conservation equations are the same as Eqs. 10, 5 and 6, respectively. Here, for a buoyant jet, the integral form of energy equation needs to be introduced (as a result of multiplying Eq. 9 by u and integrating from $y = 0$ to $y = \infty$):

$$\frac{d}{dx} \int_0^\infty \frac{\rho u^2}{2} u dy = - \int_0^\infty \tau \frac{\partial u}{\partial y} dy + \int_0^\infty g \Delta \rho u dy \quad (12)$$

Equation 12 states that the flux of kinetic energy in the jet plume is decreased by turbulence production (first term in the right hand side of Eq. 12) and increased by the work done by buoyancy (second term in the right hand side).

The Gaussian type self-similarity equations are still valid for plane buoyant jets. Using Eqs. 7 and 8, as well as the experimental results on the jet spreading rate $\frac{db}{dx}$ and on the ratio of the concentration half-width to the velocity half-width $\frac{b_c}{b}$, and after some mathematical manipulations, the solutions for a plane buoyant jet can be obtained as shown in Table 3. It is interesting to note that in Table 3, the jet centerline concentration or velocity equation are composed of two parts, corresponding to two limits (pure jet-like or pure plume-like conditions).

The characteristic length scale for a buoyant circular jet or plume is:

$$L_M = \frac{M_0^{\frac{3}{4}}}{B_0^{\frac{1}{4}}} \quad (13)$$

Similarly, for a buoyant circular jet, the integral equations of the momentum, continuity, pollutant mass (or buoyancy) conservation, and kinetic energy can be obtained. Using the Gaussian distribution for jet velocity or concentration and some experimental data, the analytical solutions for buoyant circular jets are shown in Table 3.

Effect of Boundaries on Jets and Plumes

In this section, the effect of different types of boundaries on jets and plumes will be considered, including the solid bed (wall), the free surface of ambient fluid, the coflowing or crossflowing ambient fluid, and the neighboring jets in the case of multiple jets.

Wall Jets

Wall jets are the jets discharged tangentially or at certain angles to a solid boundary (wall) (see Fig. 2). A simple case is first considered: a plane jet discharged tangentially to a smooth flat plate in deep still ambient fluid of the same kind. For turbulent plane wall jets with high Reynolds numbers ($R_0 = \frac{U_0 b_0}{\nu} = 10^4 \sim 10^5$, where ν is the kinematic viscosity of the fluid) at the slot exit, the length of the potential core will be $(6.1 \sim 6.7)b_0$ [73], which is in the same range as for simple jets. As expected, experiments show that, near the wall, there exists a thin layer (boundary layer) where the jet velocity increases from zero at the wall to a maximum velocity u_m ; above the boundary layer (named free mixing region), the jet velocity decreases from u_m to zero at some large distance y from the wall. Similarly as for simple jets, the jet width may be defined as where the jet velocity is 50% (or 37%) of u_m and $\frac{\partial u}{\partial y} < 0$ (i.e., in the free mixing region). In the boundary layer region, the boundary layer theories may be used to further divide this region into two or three sub-layers: in the sub-layer very close to the wall, the velocity distribution is linear with y ; some distance away from the wall, the velocity distribution can be described by the logarithmic law [87]. For the velocity distributions of the entire wall jet, after some distance (about $20b_0$) from the slot exit, they exhibit self-similarity [23, 99]. Verhoff [99] proposed an empirical equation which agreed well with the experimental data:

$$\frac{u}{u_m} = 1.48 \left(\frac{y}{b} \right)^{\frac{1}{7}} \left[1 - \operatorname{erf} \left(0.68 \frac{y}{b} \right) \right] \quad (14)$$

Using the equations of motion and the integral method, the following results could be obtained for plane wall jets: $u_m \propto x^{-\frac{1}{2}}$; $b \propto x$. The detailed results are listed in Table 4. To study the effect of wall roughness on wall jets, readers may refer to Rajaratnam [72], Tachie et al. [95], Dey et al. [19], and Rostamy et al. [83]. To study the jets impinging on walls, readers

Transport with Jets and Plumes of Chemicals in the Environment. Table 3 Summary of mean properties of turbulent buoyant jet

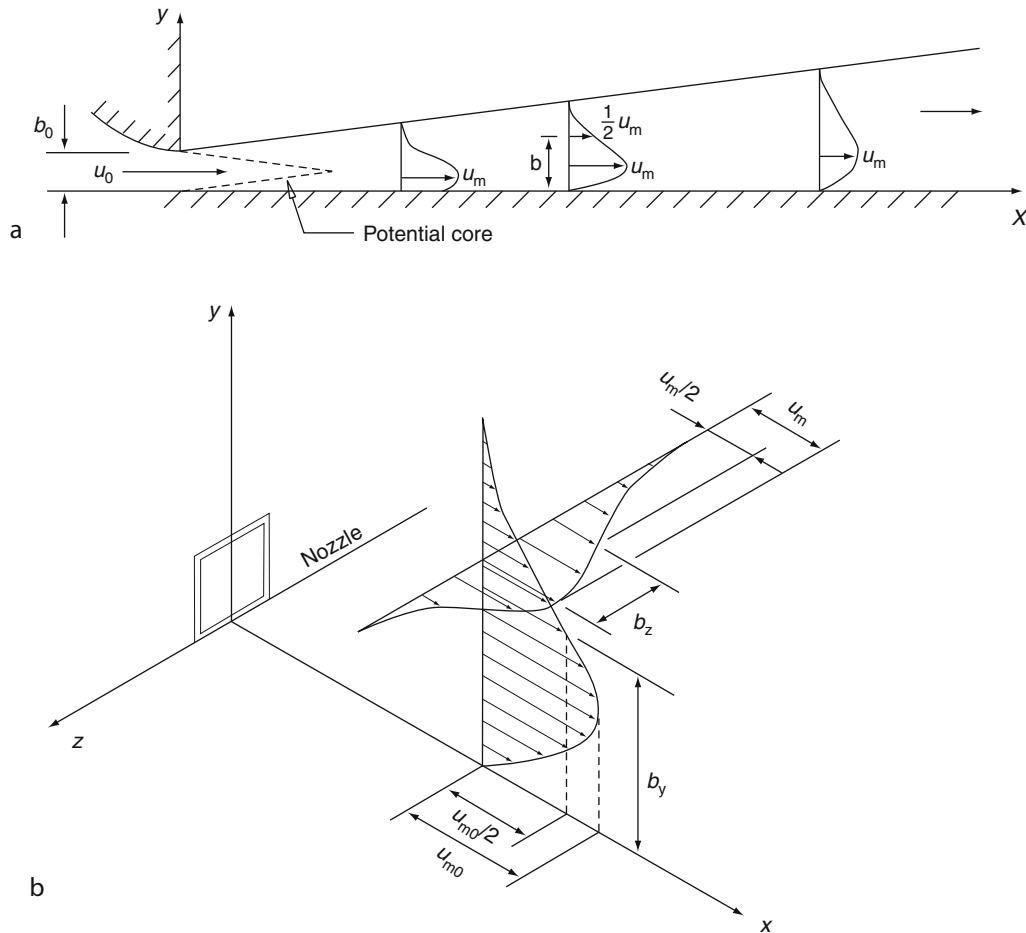
Parameter	Plane buoyant jet	Circular buoyant jet
Characteristic length for jet/plume L_M	$L_M = \frac{M_0}{B_0^{\frac{2}{3}}}$	$L_M = \frac{M_0^{\frac{3}{4}}}{B_0^{\frac{1}{2}}}$
Maximum (centerline) concentration C_m or density defect $\Delta\rho_m$	$\frac{C_m}{C_0} = \frac{\Delta\rho_m}{\Delta\rho_0} = \frac{\alpha_{11}}{\left[\frac{\alpha_{12}}{F_0^2} + \frac{\alpha_{13}}{\left(\frac{x}{b_0}\right)^{\frac{3}{2}}} \right]^{\frac{1}{3}}} \frac{1}{\left(\frac{x}{b_0}\right)}, \text{ where}$ $F_0 = \frac{U_0}{\sqrt{g \frac{\Delta\rho}{\rho_a} b_0}}, \alpha_{11} = 12.75, \alpha_{12} = 21.19, \alpha_{13} = 50.0$ in [74]	$\frac{C_m}{C_0} = \frac{\Delta\rho_m}{\Delta\rho_0} = \frac{\alpha_{11}}{\left[\frac{\alpha_{12}}{F_0^2} \left(\frac{x}{d_0}\right)^5 + \alpha_{13} \left(\frac{x}{d_0}\right)^3 \right]^{\frac{1}{3}}},$ where $\alpha_{11} = 100, \alpha_{12} = 1,920, \alpha_{13} = 6,720$ in [74]
Cross-sectional average concentration C_{avg}	$\frac{C_m}{C_{avg}} = \sqrt{1 + \frac{1}{k_2^2}} = 1.24, \text{ where } k_2 = \frac{b_c}{b} = 1.35$ in [45]	$\frac{C_m}{C_{avg}} = 1.80 \text{ in [74]}$ $\frac{C_m}{C_{avg}} = 1 + \frac{1}{k_2^2} = 1.69, \text{ where}$ $k_2 = \frac{b_c}{b} = 1.2 \text{ in [45]}$
Maximum (centerline) velocity u_m	$\frac{u_m}{U_0} = \left[\frac{\alpha_{31}}{F_0^2} + \frac{\alpha_{32}}{\left(\frac{x}{b_0}\right)^{\frac{3}{2}}} \right]^{\frac{1}{3}}, \text{ where } \alpha_{31} = 21.2,$ $\alpha_{32} = 50.7 \text{ in [74]}$	$\frac{u_m}{U_0} = \left[\frac{\alpha_{31}}{F_0^2 \left(\frac{x}{d_0}\right)} + \frac{\alpha_{32}}{\left(\frac{x}{d_0}\right)^3} \right]^{\frac{1}{3}}, \text{ where}$ $\alpha_{31} = 64.75, \alpha_{32} = 223.25 \text{ in [74]}$
Velocity half-width b	$b = \alpha_4 x$, where $\alpha_4 = 0.097$ in [74]	$b = \alpha_4 x$, where $\alpha_4 = 0.097$ in [74]
Concentration half-width b_c	$\frac{b_c}{b} = \alpha_5$, where $\alpha_5 = 1.18$ in [74]	$\frac{b_c}{b} = \alpha_5$, where $\alpha_5 = 1.16$ in [74]
Entrainment Coefficient α_e	$\alpha_e = \alpha_{ej} + \frac{\alpha_{61}(\alpha_{ep} - \alpha_{ej})}{\alpha_{61} + \frac{\alpha_{62}F_0^2}{\left(\frac{x}{b_0}\right)^{\frac{3}{2}}}}, \text{ where } \alpha_{ej} \text{ and } \alpha_{ep} \text{ are}$ $\text{the entrainment coefficients for the plane jet and plume, respectively, } \alpha_{61} = 17.1, \alpha_{62} = 41.4 \text{ in [74]}$	$\alpha_e = \alpha_{ej} + \frac{\alpha_{61}}{\alpha_{62} + \frac{\alpha_{63}F_0^2}{\left(\frac{x}{d_0}\right)^{\frac{3}{2}}}}, \text{ where}$ $\alpha_{61} = 0.44, \alpha_{62} = 20.74, \alpha_{63} = 2.38 \text{ in [74]}$

Source: [45, 74].

may refer to Beltaos and Rajaratnam [10], Rajaratnam [73], and Chan et al. [14].

For non-buoyant circular wall jets (herein termed “bluff wall jet” to include semicircular and rectangular wall jets with aspect ratio not very different from unity; the properties of bluff jets are not very different from those of circular jets), a number of studies have been conducted: square wall jets by Sforza and Herbst [88], circular wall jets by Newman et al. [61], bluff (including square, rectangular, circular, elliptic, and equilateral triangular) wall jets by Rajaratnam and

Pani [77], square wall jets by Lübecke et al. [53], and circular wall jets by Agelin-Chaab [3]. These experiments show that after some distance from the potential core, the velocity distributions both in the vertical central plane and in the horizontal plane (see Fig. 2) are self-similar. From similarity analysis on the equations of motion or from dimensional analysis, the following results can be obtained for bluff wall jets: $u_m \propto x^{-1}$; $b_y \propto x$; $b_z \propto x$ [73]. Experimental results support these predictions, and the results are listed in Table 4.



Transport with Jets and Plumes of Chemicals in the Environment. Figure 2
Schematic of (a) plane and (b) bluff wall jets

Surface Jets

A surface jet can be produced by discharging a fluid at the surface of an ambient fluid (see Fig. 3). One typical example is the surface discharge of heated water from a power plant through either an open-channel or a pipe into an ocean, a lake, or river. Rivers flowing into lakes, reservoirs, and oceans and storm water discharges into rivers may be also viewed as surface jets. In this section, our attention will be limited in the region from the end of the jet potential core to the end of the near field (where the mixing is still dominated by the jet momentum and buoyancy). The length of the near field is in the order of $100\sqrt{A_0}$, where A_0 is the cross-sectional area of the flow at discharge [74]. For the mixing in the far field (where the turbulence in rivers, lakes, or oceans

dominates further mixing), readers can refer to Fischer et al. [21] and Rutherford [85].

Non-buoyant plane surface jets in stagnant water are now considered. Equations 1–8 for plane submerged jets also work for plane surface jets. Using the integral method and some mathematical manipulations, some useful results can be obtained: $u_m \propto x^{-\frac{1}{2}}$; $b \propto x$; $c_m \propto x^{-\frac{1}{2}}$, which are in the same form as those for plane submerged jets (Table 1). Experiments on plane surface jets were conducted by Chu and Vanvari [16], Rajaratnam and Humphries [76], and others. The experimental results are listed in Table 5. The results confirm that essentially a plane surface jet is quite similar to half of the corresponding plane submerged jets, but with slightly different coefficients. For example,

Transport with Jets and Plumes of Chemicals in the Environment. Table 4 Summary of mean properties of plane wall jet and bluff wall jet

Parameter	Plane wall jet	Bluff wall jet
Maximum (centerline) concentration C_m		
Cross-sectional average concentration C_{avg}	$\frac{C_{avg}}{C_0} = \frac{\alpha_1}{\frac{x}{b_0}}$, where $\alpha_1 = 4.032$ in [73]	
Maximum (centerline) velocity u_m	$\frac{u_m}{U_0} = \frac{\alpha_3}{\sqrt{\frac{x}{b_0}}}$, where $\alpha_3 = 3.50$ in [78] & [73]	
Velocity half-width b	$b = \alpha_4 x$, where $\alpha_4 = 0.068$ in [73]; $\frac{db}{dx} = 0.073$ in [44]	$\frac{b_y}{d_0} = 0.90 + \alpha_{41} \frac{x}{h}$ $\frac{b_z}{B} = \alpha_{42} \frac{x}{B} - 1.25$ where $\alpha_{41} = 0.045$, $\alpha_{42} = 0.20$, B is the nozzle (horizontal) width in [77] & [73]; $\frac{db_y}{dx} = 0.048$ and $\frac{db_z}{dx} = 0.26$ in [44]
Concentration half-width b_c		
Entrainment Coefficient α_e	$\alpha_e = 0.035$ in [73]	

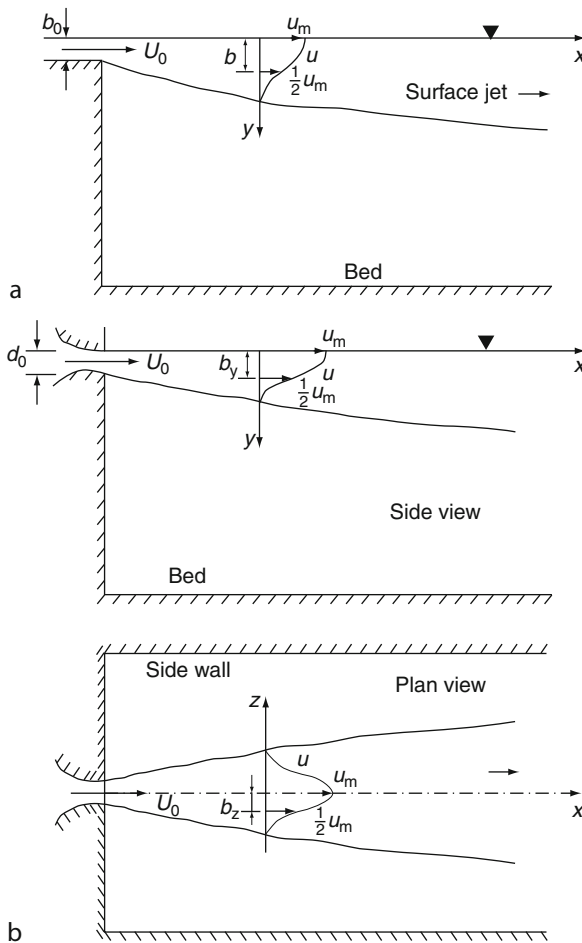
Source: [44, 73, 77, 78].

the jet spreading rate of plane surface jets $\frac{db}{dx} = 0.07$, smaller than the value of 0.10 for plane submerged jets.

Next, non-buoyant bluff surface jets in stagnant water are considered. From the experiments of Rajaratnam and Humphries [76], the Gaussian function describes the velocity distributions well both in vertical (half-Gaussian) and transverse directions, unless there is excess wave generation at the water surface (in this case u_m occurs some distance below the water surface). Using the integral method, it can be shown that: $u_m \propto x^{-1}$; $b_z \propto x$; $b_y \propto x$, which are in the same forms as for circular submerged jets. Rajaratnam and Humphries' experimental results show that the jet spreading rate in the transverse direction $\frac{db_z}{dx} = 0.09$, twice of that in the vertical direction $\frac{db_y}{dx} = 0.044$. The phenomenon of the several times faster transverse spreading has also been observed in the studies of Anthony and Willmarth [6], Gholamreza-kashi et al. [24], and Cuthbertson and Davies [18]. These studies further found that there exists a thin layer (called "surface current") at the free surface, which exhibits even faster transverse spreading

compared to that below the layer. Comparing Table 5 with Table 4, one may find that surface jets are somehow similar to wall jets, e.g., they both spread faster in the horizontal direction than in the vertical direction due to the boundary constraint in the vertical direction, and they have similar forms of jet equations.

For the surface discharges such as heated water into rivers or wastewater into the oceans, the effect of buoyancy needs to be considered. Experimental results have indicated that the behavior of buoyant surface jets is mainly controlled by three parameters: the Richardson number at the outfall, $Ri_0 = gd_0\Delta\rho_0/(\rho_a U_0^2)$; the depth (thickness) of the surface jet, d_0 (or b_0 for a plane jet); and the depth of the surface stratified layer formed at the end of the near field of the surface jet, $\overline{b_\infty}$. For a buoyant surface jet with a fixed Ri_0 , depending on the value of $\frac{d_0}{\overline{b_\infty}}$, there could be four possible hydraulic phenomena: a surface jet, a surface (density) jump at the outfall, a surface jet followed by a surface jump, or a drowned jump. Rajaratnam and Subramanyan [79] presented a graph to distinguish which of the four possibilities may happen for a plane buoyant surface jet. For the case of a pure plane



Transport with Jets and Plumes of Chemicals in the Environment. Figure 3

Schematic of (a) plane and (b) bluff surface jets

buoyant surface jet (without any jump), the experiments of Rajaratnam and Subramanyan [79] show that: initially the jet spreading rate $\frac{db}{dx}$ follows the equation of the plane non-buoyant surface jet, but after some longitudinal distance, the spreading rate slows down, and eventually the jet thickness approaches a constant. In other words, generally the buoyancy effect constrains the spreading of a plane surface jet. Their results also indicate that $\frac{u}{u_m}$ is self-similar at different x -sections; however, these self-similarities can no longer be described by the Gaussian distribution and seem to be related with Ri_0 . The results of Chu and Vanvari [16] reveal that the entrainment coefficient α_e of a plane buoyant surface jet decreases continuously with the

increase of bulk Ri (defined as $\frac{U_0 d_0 g \Delta \rho_0}{\rho_a u_m^3}$) with x ; and α_e equals to zero when Ri increases to 0.2.

For bluff buoyant surface jets, a number of experiments have shown that the jet behavior is strongly affected by Ri_0 at the outfall. For convenience, bluff surface jets may be classified into two classes, the small Ri_0 class ($Ri_0 \leq 0.1$) and the large Ri_0 class ($Ri_0 > 0.1$). From a number of experiments, the common findings for the two classes are that the vertical velocity profile $u(y)$ in the center-plane and the transverse (across the jet) velocity profile $u(z)$ just below the water surface are self-similar, and the self-similarities can be well described by half-Gaussian or Gaussian distribution. Using the similarity analysis of the simplified equations of motion, the following relations can be obtained: for the small Ri_0 class, $u_m \propto x^{-1}$, $C_m \propto x^{-1}$, $b_y \propto x$, $b_z \propto x$; and for the large Ri_0 class, $u_m \propto x^{-\frac{1}{3}}$, $C_m \propto x^{-\frac{2}{3}}$, $b_y = \text{constant}$, $b_z \propto x$. The detailed results are listed in Table 6.

Jets and Plumes in Coflow

Similarly as solid bed or free surface of ambient fluid, coflowing or crossflowing ambient fluid itself can be viewed as some sort of boundary affecting jet behaviors. When jets are discharged in the direction of flowing ambient fluids, this is the problem of jets in coflow (Fig. 4). Extensive experimental and numerical studies show that beyond the potential core, the jet concentration and the jet excess velocity relative to the ambient velocity exhibit self-similarity. The self-similarity may be described by the Gaussian distribution, exponent function, or cosine expression [45, 73]. In the following, plane jets in uniform coflow will be briefly introduced, followed by circular jets.

For coflowing plane jets, the integral form of equation of motion is:

$$\frac{d}{dx} \int_0^\infty \rho u \Delta u dy = 0 \quad (15)$$

where $\Delta u = u - U_a$, the jet excess velocity. Equation 15 states that the excess momentum flux is conserved in x direction. Using Eq. 15 and similarity analysis, Rajaratnam [73] obtained the following asymptotic relations: for the strong jet region (i.e., $\frac{\Delta u_m}{U_a} \gg 1$), $\Delta u_m \propto x^{-\frac{1}{2}}$ and $b \propto x$; and for the weak jet region (i.e., $\frac{\Delta u_m}{U_a} \ll 1$), $\Delta u_m \propto x^{-\frac{1}{2}}$ and $b \propto x^{\frac{1}{2}}$. Based on the

Transport with Jets and Plumes of Chemicals in the Environment. Table 5 Summary of mean properties of non-buoyant plane surface jet and bluff surface jet

Parameter	Non-buoyant plane surface jet	Non-buoyant bluff surface jet
Maximum (centerline) concentration C_m		
Cross-sectional average concentration C_{avg}		
Maximum (centerline) velocity u_m	$\frac{u_m}{U_0} = \frac{\alpha_3}{\sqrt{\frac{x}{b_0}}}$, where $\alpha_3 = 3.1$ in [76]	$\frac{u_m}{U_0} = \frac{\alpha_3}{x/d_0}$, where $\alpha_3 = 13$ in [24]
Velocity half-width b	$\frac{db}{dx} = \alpha_4$, where $\alpha_4 = 0.07$ in [74] & [76]	In transverse direction: $db_z/dx = \alpha_{41}$ In vertical direction: $db_y/dx = \alpha_{42}$ where $\alpha_{41} = 0.09$ in [74] & [76], 0.12 (below the free surface) and 0.22 (at the free surface) in [24]; and $\alpha_{42} = 0.044$ in [74] & [76], 0.025 in [24]
Concentration half-width b_C	$b_C/b = \alpha_5$, where $\alpha_5 = 1.15$ in [74]	
Entrainment coefficient α_e	$\alpha_e = 0.037$ in [74]	

Source: [24, 73, 74, 76].

experimental results in the literature, Rajaratnam [73] derived that:

$$\frac{\Delta u_m}{\sqrt{U_0(U_0 - U_a)}} = \frac{3.41}{\sqrt{\frac{x}{b_0}}} \quad (16)$$

$$\frac{b}{b_0} = 0.118 \frac{x}{b_0} \frac{1}{1 + \frac{0.41}{\sqrt{\alpha(\alpha-1)}} \sqrt{\frac{x}{b_0}}} \quad (17)$$

where $\alpha = \frac{U_0}{U_a}$, the ratio of jet exit velocity to ambient velocity. Note that Rajaratnam [73] also summarized other more complex forms of equations for Δu_m and b , which were derived by Patel [68] and Pande and Rajaratnam [62].

For circular jets in coflow, the integral momentum equation can be derived:

$$\frac{d}{dx} \int_0^\infty \rho 2\pi r u \Delta u dr = 0 \quad (18)$$

which says that the jet excess momentum is conserved in axial direction. Using Eq. 18 and similarity analysis on the equations of motion, Rajaratnam [73] presented the following asymptotic relations: for the strong jet

region (i.e., $\frac{\Delta u_m}{U_a} \gg 1$), $\Delta u_m \propto x^{-1}$ and $b \propto x$; for the weak jet region (i.e., $\frac{\Delta u_m}{U_a} \ll 1$), $\Delta u_m \propto x^{-\frac{2}{3}}$ and $b \propto x^{\frac{1}{3}}$. Pande and Rajaratnam [62] proposed a complex expression for Δu_m . Lee and Chu [45] also derived asymptotic solutions for circular jets in coflow: for the strong jet region, the jet solution is assumed to be the same as in stagnant water (see Table 1); for the weak jet region,

$$\frac{\Delta u_m}{U_a} = 2.14 \left(\frac{x}{l_m^*} \right)^{-\frac{2}{3}} \quad (19)$$

$$\frac{b}{l_m^*} = 0.385 \left(\frac{x}{l_m^*} \right)^{\frac{1}{3}} \quad (20)$$

where l_m^* is the excess momentum length scale defined as $\frac{M_{e0}}{U_a}$; $M_{e0} = (U_0 - U_a) \frac{U_0 \pi d_0^2}{4}$, the jet specific excess momentum at discharge.

To completely model circular jets in coflow, Lee and Chu [45] formulated an integral model based on a Lagrangian jet spreading hypothesis:

$$U^{*2} + U^* - \frac{1}{\pi B^{*2}} = 0 \quad (21)$$

Transport with Jets and Plumes of Chemicals in the Environment. Table 6 Summary of Mean properties of buoyant plane surface jet and bluff surface jet

Parameter	Buoyant plane surface jet	Buoyant bluff surface jet with $Ri_0 \leq 0.1$	Buoyant bluff surface jet with $Ri_0 > 0.1$
Maximum (centerline) concentration C_m			$\frac{C_m}{C_0} = \frac{\alpha_1}{\left(\frac{x}{\sqrt{A_0}}\right)^{\frac{2}{3}}}$, where $\alpha_1 = 2.83$ in [74]
Cross-sectional average concentration C_{avg}			
Maximum (centerline) velocity u_m	$\frac{u}{u_m}$ exhibits self-similarity, but cannot be well described by Gaussian profile, as $\frac{u}{u_m} \approx 0$ at $\frac{y}{b} \approx 1.4$ [79]	$\frac{u_m}{U_0} = \frac{\alpha_3 Ri_0^{\frac{1}{10}}}{\sqrt{\frac{x}{A_0}}}$, where $\alpha_3 = 15.3$ in [74] & [75]	$\frac{u_m}{U_0} = \frac{\alpha_3}{\left(\frac{x}{\sqrt{A_0}} - 5.0\right)^{1/3}}$, where $\alpha_3 = 1.25$ in [74]
Velocity half-width b	First $\frac{b}{b_0}$ increases linearly as non-buoyant surface jet to some point $\frac{x}{b_0}$ (the location depends on Ri_0); then the increase rate $\frac{db}{dx}$ decreases nonlinearly, finally approaches asymptotically a horizontal line [79].	$\frac{db_y}{dx} = \alpha_{41}$ $\frac{db_z}{dx} = \alpha_{42}$ where α_{41} decreases with Ri_0 ($\alpha_{41} = 0.044$ for $Ri_0 = 0$, 0.02 for $Ri_0 = 0.038$, 0 for $Ri_0 = 0.09$); α_{42} increases with Ri_0 in [74] & [75]	$\frac{b_{y*}}{d_0} = \frac{\alpha_{41}}{Ri_0^{1/8}}$ $\frac{b_z}{\sqrt{A_0}} = \alpha_{42} \left(\frac{x}{\sqrt{A_0}} + 2.0 \right)$ where b_{y*} is the average of b_y which changes slightly with y ; $\alpha_{41} = 0.29$ in [63], 0.26 in [74]; $\alpha_{42} = 0.54$ in [74]
Concentration half-width b_c		$\frac{b_{yc}}{b_y} = \alpha_{51}$ $\frac{b_{zc}}{b_z} = \alpha_{52}$ where $\alpha_{51} = 1.0$ in [74] & [75] and $\alpha_{52} = 1.15$ in [74]	$\frac{b_{yc}}{b_y} = \alpha_{51}$ $\frac{b_{zc}}{b_z} = \alpha_{52}$ where $\alpha_{51} = 1.12$ in [63] & [74]; and $\alpha_{52} = 1.9$ in [63], 1.6 in [74]
Entrainment coefficient α_e	α_e decreases from about 0.04 to 0 when Ri increases from 0 (non-buoyant plane surface jet) to 0.2 [16]		

Source: [16, 63, 74, 75, 79].

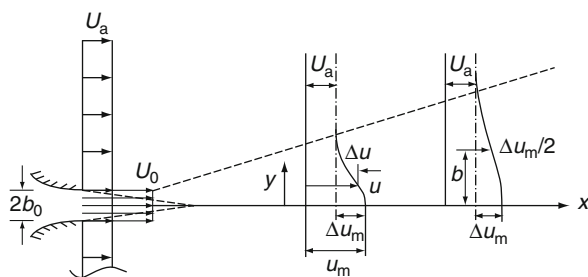
$$\frac{dB^*}{dx} = \beta_s \frac{U^*}{1 + U^*} \quad (22)$$

where $U^* = \frac{\Delta U}{U_a}$; $B^* = \frac{B}{l_m}$; ΔU and B are, respectively, the excess velocity and half of the width of the top-hat profile (instead of the Gaussian profile) of an equivalent jet, which carries the same mass flow and excess momentum flux as the actual jet; and $\beta_s = \frac{dB}{dx}$ in stagnant water. It can be proved that $\Delta U = \frac{\Delta u_m}{2}$ and $B = \sqrt{2}b$, where Δu_m and b are, respectively, the maximum excess velocity and 37% half-width for the

Gaussian profile. From Eqs. 21 and 22, ΔU and B can be solved. The actual jet centerline dilution can be obtained:

$$S_c = \frac{C_0}{C_m} = \frac{\lambda^2 \pi B^2 \left(U_a + \frac{2}{1+\lambda^2} \Delta U \right)}{2 Q_0^2} \quad (23)$$

where Q_0 and C_0 are the initial jet discharge and concentration, respectively; λ is the ratio of concentration half-width to velocity half-width using the Gaussian profile ($\lambda \approx 1.2$). The modeling results of Lee and



Transport with Jets and Plumes of Chemicals in the Environment. Figure 4

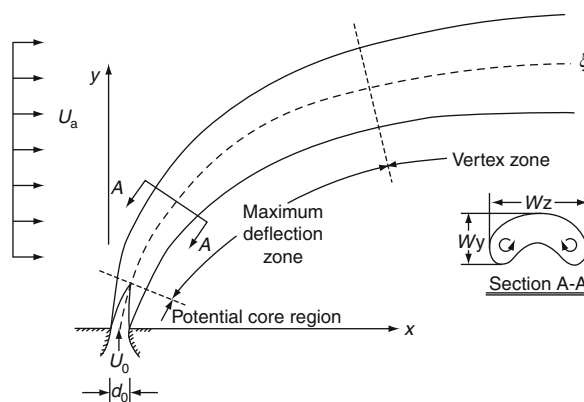
Schematic of plane jets in coflow (for circular jets, replace $2b_0$ by d_0 and y by r)

Chu [45] reveal that the centerline dilution of a circular jet in coflow is only slightly smaller than that in stagnant water; and the centerline excess velocity decays in a similar way as in stagnant water.

Jets and Plumes in Crossflow

Now consider a non-buoyant circular jet discharged at an oblique angle (not 0 or 180 degree) to a flowing ambient fluid. In fact, most outfalls or diffusers in oceans or rivers discharge effluents as jets in crossflow, as the jet (effluent) dilution can be considerably enhanced even in a weak crossflow [45, 73]. Jets in crossflow have been studied extensively by Abramovich [1], Rajaratnam [73], Fischer et al. [21], Wright [102, 103], Andreopoulos [5], Hodgson and Rajaratnam [29], Margason [54], Smith and Mungal [91], Lee and Chu [45], Huang et al. [31], Kikkert et al. [38], and others. According to these studies, the evolution of jets in crossflow can be divided into three regions: the potential core region, the maximum deflection region, and the vortex region (see Fig. 5). The length of the potential core has been found to be mostly controlled by the relative strength of the jet compared to the crossflow ($\alpha = \frac{U_0}{U_a}$), and typically in the range of $2-6d_0$ which is smaller than that of a free jet [70].

Beyond the potential core region, the jet would be largely deflected due to the stagnation pressure exerted by the free stream and the entrainment of ambient fluid (and thus horizontal momentum). Jet deflection probably is the most distinct feature in crossflow. After the maximum deflection region, the jet would be gradually parallel to the direction of ambient flow. Laboratory



Transport with Jets and Plumes of Chemicals in the Environment. Figure 5

Schematic of jets in crossflow

experiments [1, 29, 45] and numerical simulations [45] have found that after some distance beyond the potential core, the jet cross section would be like a kidney shape with a pair of two counter-rotating vortices (see Fig. 5). The vortex pair significantly entrains ambient fluid in the form of tornado vortices into the jet [45], which explains the considerable enhancement of jet dilution in crossflow. The concentrations at the centers of the two vortices have been found to be about 1.1–1.6 times of the jet centerline concentration [29, 45].

For non-buoyant jets in crossflow, it is common to analyze them in three regions: the momentum dominated near field (MDNF), the momentum dominated far field (MDFF), and the transition between MDNF and MDFF. If the jet trajectory $\xi \ll l_m$, where l_m is defined as:

$$l_m = \frac{M_{v0}^{1/2}}{U_a} \quad (24)$$

and M_{v0} is the vertical momentum at the exit, then the jet is in MDNF, where the effect of jet momentum is much stronger than that of the ambient crossflow. In MDNF, the classic equations for jets in stagnant ambient fluid are approximately valid. If $\xi \gg l_m$, the jet is in MDFF, where the effect of ambient crossflow is dominant over the jet momentum. In MDFF, the jet properties can be studied with physical and numerical models. Fischer et al. [21] used dimensional analysis to find the asymptotic formulas for MDNF and MDFF.

Lee and Chu [45] studied the jets in MDFF based on the analogy to advected line puffs. Using length-scale analysis and numerical models, they proposed the formulas which can represent satisfactorily the experimental results:

$$\frac{y_c}{l_m} = 1.56 \left(\frac{x}{l_m} \right)^{\frac{1}{3}} \quad (25)$$

$$b_{vc} = 0.28 y_c \quad (26)$$

$$S_c = 0.46 \frac{U_a y_c^2}{Q_0} \quad (27)$$

where y_c is the vertical location of the centerline concentration, b_{vc} is the vertical centerline half-width defined by 37% of the centerline concentration and S_c is the centerline dilution.

The most common jet discharge angle in crossflow is 90 degree, i.e., jets are discharged at right angle to the ambient crossflow. Rajaratnam [73] summarized the early studies in 1950s to 1970s that mostly focused on jet trajectories. Hodgson and Rajaratnam [29] conducted detailed laboratory experiments on circular jets at right angle to crossflow and proposed the following equations:

$$S_c = 1.09 \left(\frac{\alpha x}{d_0} \right)^{0.56} \quad (28)$$

$$\frac{y_c}{\alpha d_0} = 1.46 \left(\frac{x}{\alpha d_0} \right)^{0.26} \quad (29)$$

$$\frac{W_z}{\alpha d_0} = 1.20 \left(\frac{x}{\alpha d_0} \right)^{0.29} \quad (30)$$

$$\frac{W_y}{\alpha d_0} = 0.78 \left(\frac{x}{\alpha d_0} \right)^{0.37} \quad (31)$$

where W_z and W_y are the jet width and thickness (see Fig. 5). Equations 28–31 have also been validated by a field experiment in the Lesser Slave River, Canada. Hodgson and Rajaratnam's equations are mainly derived based on the experiments conducted in the range of $\frac{x}{\alpha d_0} = 1 \sim 1,000$. It is interesting to note that Hodgson and Rajaratnam's expressions fit the experimental data satisfactorily both in MDNF, MDFF, and the transition between the two.

Now consider a circular plume in crossflow. Similarly as l_m , a length scale l_b needs to be defined to

compare the relative strength of plume buoyancy with crossflow:

$$l_b = \frac{B_0}{U_a^3} \quad (32)$$

If the jet trajectory $\xi \ll l_b$, then the plume is in the buoyancy dominated near field (BDNF) where the effect of buoyancy is dominant over crossflow; if $\xi \gg l_b$, then plume is in the buoyancy dominated far field (BDFF) where the effect of crossflow is more pronounced than the buoyancy. In BDNF, the plume is essentially vertical and only slightly advected, thus the equations for plumes in stagnant fluid are approximately valid. Similarly as jets in crossflow, after some distance from the nozzle, the plume cross section will become a kidney shape that is made up of a vortex pair, and the concentration at the centers of the vortices have been found to be 1.4–1.7 times of the plume centerline concentration. In BDFF, the plume bends over and finally approaches the ambient flow direction. The analysis on the plume properties in BDFF relies on experiments or numerical models. Based on the equations of motion and the use of similarity solutions, Fischer et al. [21] derived asymptotic formulas for the BDNF and BDFF. As the plume in the BDFF behaves similarly as the advected line thermal, Lee and Chu [45] used numerical models to obtain the plume characteristics. The predictions are comparable to experimental results. The formulas Lee and Chu derived are:

$$\frac{y_c}{l_b} = 1.3 \left(\frac{x}{l_b} \right)^{\frac{2}{3}} \quad (33)$$

$$b_{vc} = 0.4 y_c \quad (34)$$

$$S_c = 0.46 \frac{U_a y_c^2}{Q_0} \quad (35)$$

Note Eq. 35 for BDFF is exactly in the same form as the equation for MDFF.

Now consider the case of a circular buoyant jet in crossflow. In this case, the relative strengths of buoyancy, momentum, and crossflow need to be considered. If $l_b \ll l_m$, then the jet would in sequence experience in MDNF, MDFF, and BDFF; if $l_b \gg l_m$, then the sequence would be MDNF, BDNF, and BDFF [21, 43, 45]. Equations should be selected carefully according to the studied location of the jet (e.g., in MDNF or MDFF or BDNF or BDFF). The transitions between

MDNF and MDFF and between BDNF and BDFF are better treated using numerical models [45].

For jets directed at an oblique angle to crossflow, the reader may refer to Platten and Keffer [69] and Kikkert et al. [38]. For plane jets and plume in crossflow, the reader may refer to Girshovich [25], Jones and Wille [36], Kalita et al. [37], and Huang et al. [31].

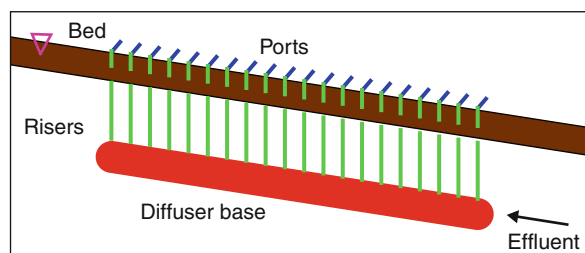
Multiple Jets

Effluents may be discharged via single port outfalls or multiport diffusers. Multiport diffusers are commonly used given their fast mixing and diluting ability and thus less adverse impacts on the environment. The jets issuing from the ports of a multiport diffuser are usually viewed as multiple jets. The characteristics of multiple jets are primarily determined by the arrangements of multiport diffusers. Generally, multiport diffusers can be classified into three categories: unidirectional diffuser (where net horizontal momentum flux is imparted perpendicular to diffuser line), staged diffuser (where net horizontal momentum flux is imparted parallel to the diffuser line), and alternating diffuser (where no net horizontal momentum flux is imparted) [20]. Figure 6 illustrates one typical example of a unidirectional diffuser. It is expected that different types of diffusers have significantly different jet mixing and spreading properties.

Studies on multiple jets (or diffusers) have been reported extensively in the past decades ([2, 21, 32, 34, 35, 39, 40, 45, 46, 94, 96, 100, 107], etc.). Most of these studies focused on the deep water ambient condition (e.g., in oceans and lakes), and limited studies

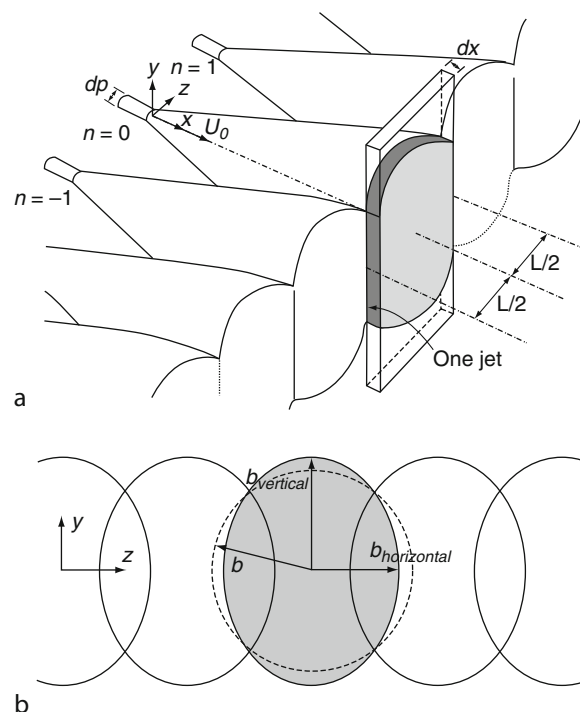
deal with the shallow water condition (e.g., in rivers). In this section, only the deep water condition will be considered. Theoretically, multiple jets in sequence experience: the individual free jet zone (where jets has no effect on each other), the jet merging zone (where the interaction between jets are strong), and the two-dimensional zone thereafter [66, 107] (see Fig. 7). In practice, multiple jets are usually simplified as one line momentum source, neglecting the interactions between individual jets that are complex and not well understood [35, 46].

In the previous sections, some basic characteristics of three-dimensional free jets and two-dimensional (plane) jets have been introduced; hence, in the following, the studies on the merging process of unidirectional non-buoyant circular jets will be briefly summarized. As yet, only limited studies on jet merging have been reported [30, 40, 64, 71, 100]. To calculate the concentration or velocity field in the jet merging



Transport with Jets and Plumes of Chemicals in the Environment. Figure 6

An example of a unidirectional diffuser



Transport with Jets and Plumes of Chemicals in the Environment. Figure 7

(a) Schematic of multiple merging jets, with the indication of (b) jet cross-section deformation (Modified from [100])

process, the most widely used method is superposition. However, as the momentum equation is nonlinear, simple superposition of individual jets would overestimate the jet velocity. Knystautas [40] studied the velocity field of merging jets in still ambient fluid and showed that the jet velocity can be modeled by superposing the momentum (u^2) of individual jets (based on Reichardt's hypothesis). Hodgson et al. [30] reported the following equation for the velocity field of the merging jets in still water:

$$u^2 = \sum_{i=-n}^{i=+n} u_i^2 = u_m^2 \sum_{i=-n}^{i=+n} \exp \left[-150 \left(\frac{z - iL}{x} \right)^2 \right] \quad (36)$$

where $(2n + 1)$ is the number of jets; the centerline velocity of each individual jet $u_m = \frac{6.13 U_0 d_0}{x}$; L is the distance between the centers of neighboring jets; z is the transverse distance from the central jet axis. Hodgson et al. [30] extended the Reichardt's hypothesis on lateral transport of momentum to the lateral transport of pollutant, and showed that uC is additive. After solving the velocity field from Eq. 36, the pollutant concentration field can be obtained from Eq. 37:

$$uC = \sum_{i=-n}^{i=+n} u_i C_i = u_m C_m \sum_{i=-n}^{i=+n} \exp \left[-130 \left(\frac{z - iL}{x} \right)^2 \right] \quad (37)$$

where the centerline concentration of each individual jet $C_m = \frac{5.34 C_0 d_0}{x}$. Hodgson's experimental results validated the use of Eqs. 36 and 37.

Hodgson et al. [30] revealed some basic physics in the jet merging, where the jet spreading rate $\frac{db}{dx}$ and the ratio of the jet concentration half-width to the velocity half-width $\frac{b_C}{b}$ are both assumed to be constant during merging. Wang and Davidson [100] developed a similar model for jet merging in stagnant ambient fluid, but allows the change of $\frac{db}{dx}$ and $\frac{b_C}{b}$ during merging, as well as the change of these parameters in horizontal (the jet merging) plane and vertical (the free entrainment) plane. Theoretical analysis and experimental data indicate that the jet merging in stagnant ambient fluid occurs at $4.5 < \frac{x}{L} < 12$. Note that in Wang and Davidson [100], the start of merging refers to the location where the jet interaction begins to influence the bulk properties of the central jet, which

is beyond the location where the physical jet boundaries start intersecting. By studying their experimental data and that of Knystautas [40], Wang and Davidson [100] found that, during the jet merging, the jet spreading rate $\frac{db}{dx}$ (or $\frac{db_C}{dx}$) increases by 30% in the vertical plane, while it decreases by a similar amount in the horizontal plane; the ratio of velocity half-width in the vertical plane to that in the horizontal plane $\frac{b_{vertical}}{b_{horizontal}}$ increases from 1 to 1.5 during merging; and the ratio of concentration half-widths $\frac{b_{C,vertical}}{b_{C,horizontal}}$ increases from 1 to 1.8. Obviously, the jet merging process constrains the jet spreading and thus dilution in the jet merging plane, while accelerates them in the free entrainment plane. This phenomenon is similar to the boundary effects found in wall jets or surface jets.

In recent years, researchers started to study the jet merging in coflow [64, 71]. Pun et al. [71] developed a multiple-point hybrid model for merging jets in coflow, which combines a length-scale model and an Eulerian-integral model. The model allows multiple transition points for each parameter (jet velocity, spread and dilution), instead of a single transition point for all these parameters. The multiple-point hybrid model is shown to be able to significantly reduce transition errors during merging compared to the single-point model, and predicts favorable results compared to the integral solution. Pani et al. [64] developed a model based on Reichardt's hypothesis for multiple coflowing jets. Instead of momentum (u^2) is additive in stagnant water, Pani et al. showed that the excess momentum ($u\Delta u$) is additive in coflow and follows Gaussian distribution. Using the method of superposition and a generalized spreading hypothesis, Pani et al. presented the equations for predicting the velocity field and centerline dilution downstream of multiple circular jets in coflow, which appeared to agree with the experimental data.

Multiphase Jets and Plumes

In this section, two kind of multiphase jets and plumes will be introduced: bubbly jets and plumes and slurry jets and plumes, which have wide engineering applications.

Bubbly Jets and Plumes

Bubbly jets are produced by injecting gas-liquid mixtures into liquids, while bubble plumes are

produced by injecting gases into liquids. Bubble plumes and bubbly jets are widely used to achieve artificial aeration, circulation and mixing in confined reactors, aeration tanks, polluted water bodies, ice-covered rivers, and deep stratified lakes and reservoirs [48, 86, 101, 104]. Such kind of gas-liquid two-phase flow is also common in some hydraulic structures, e.g., the super-gas saturation downstream of hydro-power dams. So far, most of the early studies were conducted in confined setups, where the sizes and geometry of the setup further complicates the characteristics of bubbly jets and plumes [51]. In this section, the studies in stagnant water of relatively large setup will be introduced and then the case with flowing ambient fluid will be considered.

For two-phase flows, the dissolving of the gas phase into the liquid phase can be derived from Fick's law of diffusion as [60]:

$$\frac{dC}{dt} = K_L a (C_s - C) \quad (38)$$

where C is the dissolved gas (e.g., oxygen) concentration in the liquid, C_s is the saturation dissolved gas concentration, t is the time, K_L is the mass transfer coefficient, and a is the gas-liquid interfacial area per unit liquid volume (also named the specific interfacial area). From Eq. 38, the gas transfer rate is mainly controlled by K_L and a , which differ significantly in different setups. Previous studies [9, 59] have shown that these two parameters are greatly influenced by bubble size. Bubble size depends on a number of factors: nozzle sizes and types, initial gas volume fractions, the solubility and mass transfer ability of the gas, turbulence intensity and flow structure of the ambient liquid, impurities and surfactants in the ambient liquid, etc. [17, 49–52]. Lima Neto et al. [50] proposed a criterion to judge the sizes and shapes of the bubbles produced by injecting a mixture of air and water into water: if the nozzle Reynolds number $Re = \frac{U_{w0} d_0}{\nu_w} < 8000$ (where U_{w0} is the superficial water velocity based on the water discharge at the nozzle and nozzle diameter d_0), then large and irregular bubbles will be produced; if $Re \geq 8,000$, smaller and uniform bubbles will be produced. A decrease in gas discharge or an increase in liquid discharge will decrease the bubble size [50, 52, 98].

Now consider the vertical injection of a pure gas into a pure stagnant liquid. The bubbles produced at the orifice will coalesce/breakup and rise, inducing ambient liquid entrained into the bubble core and the dissolving of the gas into the ambient. Lima Neto et al. [49] studied air injection into still water with six different nozzles (single orifice, multiple orifices, and airstone), and found that the water entrained into the bubble core under different initial air discharges Q_a and nozzle types can be described as a function of Q_a and vertical distance from the nozzles.

For a vertical bubbly jet in stagnant liquid, Milgram [56], Brevik and Kristainsen [12], and Lima Neto et al. [50] reported that the bubble area typically only occupies 50–90% of the bubbly jet in the radial direction. Lima Neto et al. [50] studied bubbly jets produced by injecting a mixture of air-water into stagnant water and found that the more uniform and smaller the bubble sizes, the wider the bubble core can spread in the radial direction; within the bubbly jets, the radial distributions of the time-averaged bubble concentration (void fraction) and water velocity of the mean flow can be well described by Gaussian distributions, similarly as for single-phase jets or plumes; and db/dx of the bubbly jets is close to that of the pure water jet.

Although the existence of bubbles seems not to change the Gaussian profiles, the entrainment of the ambient into the bubbly jets is significantly enhanced. Milgram [56], Socolofsky and Adams [92], Brevik and Kristainsen [12], and Lima Neto et al. [50] reported the entrainment coefficient of bubbly jets is in the range of 0.03–0.15, much larger than the values of pure jets or plumes. The additional entrainment probably associates with the bubble wakes [47] and additional liquid turbulence caused by interactions of the bubbles and their wakes [50]. At a specific height of the centerline of a bubbly jet, Lima Neto et al. [50] compared the liquid volume flux Q_w of the bubbly jet with that of a pure water jet with the same nozzle diameter and water flow rate:

$$\frac{Q_w}{Q_{w(C_0=0)}} = 1 + 6.426 \times 10^6 \left(\frac{C_0}{Re} \right)^{\frac{3}{2}} \quad (39)$$

where the initial gas volume fraction $C_0 = \frac{Q_{a0}}{(Q_{a0} + Q_{w0})}$; Q_{a0} and Q_{w0} are the initial volumetric flow rates of air and water, respectively. It is expected that the average

dilution of a bubbly jet is larger than that of a pure jet due to the additional entrainment of ambient water as shown in the second term of the the right hand side of Eq. 39.

For a bubbly jet injected horizontally into a stagnant liquid, there are only limited experimental studies [22, 52, 57, 98]. Lima Neto et al. [52] studied the injection of air-water mixtures into a water tank of relatively large size and reported that: first, the bubbly jet comes out of the nozzle as a whole quasi-horizontal bubbly jet where bubble breakup/coalescence occurs and only a few bubbles escape from the bubbly jet and rise vertically due to buoyancy; then, there follows a separation zone where the quasi-vertical bubble plume partially separates from the water jet (when the initial gas volume fraction $C_0 < 0.15$, the bubble plume will completely separate from the water jet); finally, the bubble plume continues rising and the water jet impinges the water surface and becomes surface jet. In Lima Neto et al. [52], the length and width of the bubble plume as well as the centerline trajectories of the bubble plume and water jet were also proposed in dimensionless forms.

For bubbly jets, bubble properties and mean liquid flow could be non-dimensionalized as functions of the initial gas volume fraction and nozzle Reynolds number for the vertical injection or as functions of the initial gas volume fraction and nozzle densimetric Froude number for the horizontal injection [50, 52]. The variation of bubble properties and mean liquid flow along the jet centerline and across the jet needs further studies. For bubble plume modeling, the reader may refer to Bravo et al. [11] and others.

Crossflow will exert significant effects on bubbly jets or bubble plumes, of which the most distinguishing one is the possibility of separation of bubble plumes (named generally as dispersed phases) from the entrained ambient fluids (continuous phases) [92]. So far, very limited studies have been reported on bubbly jets or bubble plumes in crossflow. Socolofsky and Adams [92] conducted laboratory experiments on bubble plumes produced by injecting air, air and oil, as well as air and alcohol in uniform crossflow. In weak crossflow, the separation between the bubble plumes and the entrained fluid does not occur before the plumes reach the surface. While in strong crossflow, the separation is significant and the separation height h_s can be given by an empirical relation:

$$h_s = \frac{5.1 B_0}{(U_a u_s^{2.4})^{0.88}} \quad (40)$$

where U_a is the horizontal crossflow velocity; u_s is the bubble slip (terminal rise) velocity; $B_0 = Q_0 g'$; Q_0 is the discharge of the plume fluid; $g' = \frac{(\rho_a - \rho_p)g}{\rho_a}$; ρ_a is the ambient density of water; and ρ_p is the density of the plume fluid. Socolofsky and Adams also reported that before h_s , the plumes can be treated as single-phase plumes; after h_s , the bubble plumes follow the trajectory of the vector sum of u_s and U_a (i.e., the bubble plumes rise in a linear line), and the separated entrained fluid behaves like a momentum jet (the momentum is gained by the acceleration of bubbles before the separation).

Slurry Jets and Plumes

Slurry jets are produced by injecting a mixture of liquid phase and solid phase (such as sand or clay particles). Slurry jets have wide applications in pumping industrial (e.g., mining or petroleum) tailings into settling tanks, dredging and land reclamation, discharging storm water and industrial waters that have solid particles, etc. A number of experimental and numerical studies in this area have been reported [7, 13, 26, 33, 55, 67, 90]. Usually, two injection ways were used, vertically upward and vertically downward. In the following, the main focus will be on the vertically downward injection of slurry jets into stagnant ambient fluids. Compared to single-phase jets, the adding of the solid phase will change the properties of the flow [89].

Previous experiments have indicated that the velocity and concentration of the solid phase across slurry jets follow self-similar Gaussian distributions [26, 33, 90]. In the slurry jets with dilute solid particles, Jiang et al. [33] reported that the velocity and concentration of the liquid phase also exhibit self-similar Gaussian profiles. However, this may not be valid for the slurry jets with high concentration solid particles.

The spreading of the solid phase has been found to increase linearly along the axial direction [13, 26, 55]. Brush [13] reported that the spreading rate of the velocity of the solid phase $\frac{db_s}{dx}$ depended on the particle size. Mazurek et al. [55] confirmed this by photographic measurements on sand jets, and further generalized the spreading as a function of the initial

densimetric Froude number of the solid particle. Recently, Hall et al. [26] conducted detailed experiments on pure sand jets and sand-water slurry jets in stagnant water. With the densimetric Froude number at the nozzle exit ($F_0 = \frac{U_0}{\sqrt{\frac{g d_0 (\rho_s - \rho_w)}{\rho_w}}}$ where U_0 is the initial

velocity of the sand particle from a nozzle of diameter d_0 and ρ_s is sand density) in the range of $2 \sim 6$, $\frac{db_s}{dx}$ measured $0.087 \sim 0.109$, not very different from that of single-phase jets in stagnant fluids. The difference lies in the ratio of concentration to velocity spreading rates of the solid phase ($\frac{b_{sC}}{b_s}$). Hall et al. [26] reported that, for sand jets and slurry jets with high F_0 ($F_0 \approx 6$), $\frac{b_{sC}}{b_s} = 0.86 \sim 0.92$, which means the sand concentration spreads slower than the velocity; and for slurry jets with low F_0 ($F_0 \approx 2$), $\frac{b_{sC}}{b_s} \approx 1.0$ which means they have almost equal spreading rates. This finding is contrary to the classic single-phase jet theory, which states that the concentration scale spreads faster than the velocity scale ($\frac{b_C}{b} \approx 1.2$).

Similar as single-phase jets, along the axial direction of slurry jets, the velocity and concentration of both solid phase and liquid phase decay rapidly. According to Hall et al. [26], beyond the potential core (about $2.9 d_0 F_0^{\frac{2}{3}}$), the axial concentration of the solid phase can be well described by:

$$\frac{C_m}{C_0} = \frac{17.12}{\left(\frac{x}{d_0 F_0^{\frac{2}{3}}}\right)^{\frac{5}{3}} + 11.39} \quad (41)$$

In Eq. 41, the $-5/3$ power relation is very similar to that of single-phase plume (Table 2), as Eq. 41 was established in the region at $x > 5L_M$ where the buoyant slurry jet behaved like a slurry plume [65]. Similar $-5/3$ power relation was also built for the sand concentration in sand jets. For both sand jets and slurry jets, the axial velocity of the solid phase was found to decrease rapidly and then reach a final plateau region. Generally, before the plateau region, the axial velocity of the solid phase in slurry jets can be well represented by:

$$\frac{u_m F_0}{U_0} = \frac{1.63}{\left(\frac{x}{d_0 F_0^{\frac{2}{3}}}\right)^{\frac{1}{3}} + 0.56} \quad (42)$$

Similar as Eq. 42, for sand jets, before reaching the velocity plateau region, the axial sand velocity was also found to follow the $-1/3$ power relation, which is very similar to that of single-phase plumes (Table 2). The terminal (settling) centerline velocity of the solid phase was found in the range of $0.32 \sim 0.43$ m/s, which is larger than the settling velocity of 0.033 m/s for individual solid particles. The larger terminal velocity probably can be attributed to the interactions between solid particles, i.e., the wake of previous solid particles tends to decrease the drag forces for the following particles.

For the studies dealing with particle interactions, the reader may refer to Lain and Garcia [42], Tamburello and Amitay [93], and Yan et al. [106]. For the effect of solid particle size on velocity distribution, concentration profile and turbulent properties, the reader may refer to Azimi et al. [8]. For vertically upward slurry jets, the reader may refer to Jiang et al. [33].

Future Directions

Some of the basic characteristics of a variety of jets and plumes have been reviewed, including simple jets and plumes, buoyant jets, surface jets, wall jets, jets and plumes in coflow and crossflow, multiple jets, bubbly jets and plumes, and slurry jets and plumes. The turbulence and turbulence structures in these flows have not been discussed. Interested readers may refer to the works of Heskestad [27], Wygnanski and Fielder [105], and Launder and Rodi [44]. Also, the behaviors of jets and plumes in stratified environment have not been considered. The readers may refer to the works of Morton et al. [58], Turner [97], and Roberts et al. [80, 81]. To facilitate the applications of jets and plumes theories, software packages have been developed. The USEPA-supported CORMIX is perhaps the most commonly used expert system for dealing with environmental problems involving jets and plumes. Other models are VISJET and Visual Plume. In addition to the simple jets and plumes, all the other varieties of jets and plumes are still currently under active studies. This constitutes the general tone for the future research. Herein, some of key areas are listed as follows:

- Physics and models of bubbly jets, and their application in aeration and mixing of ponds, lakes, and wastewater treatment plants

- Physics and models of slurry jets, and their industrial applications
- Physics and models of three phase jets and plumes, especially oil-water-gas plumes produced by oil spills in oceans
- Physics and models of jets and plumes in stratified environment
- Development, improvement, and validation of computational fluid dynamics (CFD) modeling (such as direct numerical modeling (DNS); large eddy simulation (LES); k-epsilon modeling; and others).

Bibliography

Primary Literature

- Abramovich GN (1963) The theory of turbulent jets. MIT Press, Cambridge, MA, 671 p
- Adams EE (1982) Dilution analysis for unidirectional diffusers. ASCE J Hydraul Div 108(HY3):327–342
- Agelin-Chaab M (2010) Experimental study of three-dimensional offset jets and wall jets. Ph.D. thesis, Department of Mechanical and Manufacturing Engineering, University of Manitoba, Winnipeg
- Albertson ML, Dai YB, Jensen RA, Rouse H (1950) Diffusion of submerged jets. Trans ASCE 115:639–697
- Andreopoulos J (1985) On the structure of jets in a crossflow. J Fluid Mech 157:163–197
- Anthony DG, Willmarth WW (1992) Turbulence measurements in a round jet beneath a free surface. J Fluid Mech 243:699–720
- Awaya Y, Fujisaki K, Matsunaga K (1985) Transition in the behavior of sediment laden vertical buoyant jet. J Hydrosoci Hydr Eng 3(1):63–74
- Azimi AH, Zhu DZ, Rajaratnam N (2010) Effect of particle size on the characteristics of sand jets in water. J Eng Mech (in press)
- Barnhart EL (1969) Transfer of oxygen in aqueous solutions. J Sanit Eng Div 95(3):645–661
- Beltaos S, Rajaratnam N (1972) Plane turbulent impinging jets. J Hydr Res 11:29–59
- Bravo HR, Gulliver JS, Hondzo M (2007) Development of a commercial code-based two-fluid model for bubble plumes. Environ Modell Softw 22(4):536–547
- Brevik I, Kristiansen Ø (2002) The flow in and around air bubble plumes. Int J Multiph Flow 28:617–634
- Brush LMJ (1962) Exploratory study of sediment diffusion. J Geophys Res 67(4):1427–1433
- Chan T, Zhou Y, Liu M, Leung C (2003) Mean flow and turbulent measurements of the impingement wall jet on a semi-circular convex surface. Exp Fluids 34(1):140–149
- Chen CJ, Rodi W (1980) Vertical turbulent buoyant jets – a review of experimental data. Pergamon, Oxford. 83 p
- Chu VH, Vanvari MR (1976) Experimental study of turbulent stratified shearing flow. J Hydraul Div ASCE 102(6):691–706
- Clift R, Grace JR, Weber ME (1978) Bubbles, drops, and particles. Academic, New York, 380 p
- Cuthbertson AJS, Davies PA (2008) Deposition from particle-laden, round, turbulent, horizontal, buoyant jets in stationary and coflowing receiving fluids. J Hydraul Eng 134(4):390–402
- Dey S, Nath TK, Bose SK (2010) Fully rough submerged plane wall-jets. J Hydro Environ Res 4(4):301–316
- Doneker RL, Jirka GH (2007) CORMIX user manual – a hydrodynamic mixing zone model and decision support system for pollutant discharges into surface waters. U.S. EPA-823-K-07-001, pp 1–236
- Fischer HB, List EJ, Imberger J, Brooks NH (1979) Mixing in inland and coastal waters. Academic, New York, 483 p
- Fonade C, Doubrovine N, Maranges C, Morchain J (2001) Influence of a transverse flowrate on the oxygen transfer performance in heterogeneous aeration: case of hydro-ejectors. Water Res 35(14):3429–3435
- Förthmann E (1936) Turbulent jet expansion. English translation N.A.C.A. TM-789. (Original paper in German, 1934. Ing. Archiv., 5)
- Gholamreza-kashi S, Martinuzzi RJ, Baddour RE (2007) Mean flow field of a nonbuoyant rectangular surface jet. J Hydraul Eng 133(2):234–239
- Girshovich TA (1966) Theoretical and experimental study of a plane turbulent jet in a crossflow. Fluid Dyn 9:84–86
- Hall N, Elenany M, Zhu DZ, Rajaratnam N (2010) Experimental study of sand and slurry jets in water. J Hydraul Eng 136(10):727–738
- Heskestad G (1965) Hot-wire measurements in a plane turbulent jet. Trans ASME J Appl Mech 32:1–14
- Hinze JO, Van Der Hegge Zijnen BG (1949) Transfer of heat and matter in the turbulent mixing zone of an axially symmetrical jet. Appl Sci Res A1:435–461
- Hodgson JE, Rajaratnam N (1992) An experimental study of jet dilution in crossflows. Can J Civ Eng 19:733–743
- Hodgson JE, Moawad AK, Rajaratnam N (1999) Concentration field of multiple circular turbulent jets. J Hydr Res 37(2):249–256
- Huang JF, Davidson MJ, Nokes RI (2005) Two-dimensional and line jets in a weak cross-flow. J Hydr Res 43(4):390–398
- Huang H, Fergen RE, Rroni JR, Tsai JJ (1996) Probabilistic analysis of ocean outfall mixing zones. J Environ Eng 122(5):359–367
- Jiang JS, Law AWK, Cheng NS (2005) Two-phase analysis of vertical sediment-laden jets. J Eng Mech 131(3):308–318
- Jirka GH, Akar PJ (1991) Hydrodynamic classification of submerged multiport-diffuser discharges. J Hydr Eng 117(9):1113–1128
- Jirka GH, Harleman DRF (1973) The mechanics of submerged multiport diffuser for buoyant discharges in shallow water. R.M. Parsons Laboratory Report, Massachusetts Institute of Technology, Boston

36. Jones WP, Wille M (1996) Large-eddy simulation of a plane jet in a cross-flow. *Int J Heat Fluid Flow* 17:296–306
37. Kalita K, Dewan A, Dass AK (2002) Predication of turbulent plane jet in crossflow. *Numer Heat Transfer A* 41:101–111
38. Kikkert GA, Davidson MJ, Nokes RI (2009) A jet at an oblique angle to a cross-flow. *J Hydro Env Res* 3(2):69–76
39. Kim DG, Seo IW (2000) Modeling the mixing of heated water discharged from a submerged multiport diffuser. *J Hydraul Res* 38(4):259–269
40. Knystautas R (1964) The turbulent jet from a series of holes in Line. *Aeronaut Q* 15:1–28
41. Kotsivos NE, List EJ (1977) Plane turbulent buoyant jet. Part I. Integral properties. *J Fluid Mech* 81:25–44
42. Lain S, Garcia J (2006) Study of four-way coupling on turbulent particle-laden jet flows. *Chem Eng Sci* 61:6775–6785
43. Lam KM, Lee WY, Chan CHC, Lee JHW (2006) Global behaviors of a round buoyant jet in a counterflow. *J Hydr Eng* 132(6):589–604
44. Launder BE, Rodi W (1983) The turbulent wall jet - measurement and modeling. *Ann Rev Fluid Mech* 15:429–459
45. Lee JHW, Chu VH (2003) Turbulent jets and plumes – a lagrangian approach. Kluwer, Norwell, 390 p
46. Lee JHW, Jirka GH (1980) Multiport diffuser as line source of momentum in shallow water. *Water Resour Res* 16(4):695–708
47. Leitch AM, Baines WD (1989) Liquid volume flux in a weak bubble plume. *J Fluid Mech* 205:77–98
48. Lima Neto IE, Zhu DZ, Rajaratnam N, Yu T, Spafford M, McEachern P (2007) Dissolved oxygen downstream of an effluent outfall in an ice-cover river: natural and artificial aeration. *J Env Eng* 133(11):1051–1060
49. Lima Neto IE, Zhu DZ, Rajaratnam N (2008) Air injection in water with different nozzles. *J Environ Eng* 134(4):283–294
50. Lima Neto IE, Zhu DZ, Rajaratnam N (2008) Bubbly jets in stagnant water. *Int J Multiph Flow* 34(12):1130–1141
51. Lima Neto IE, Zhu DZ, Rajaratnam N (2008) Effect of tank size and geometry on the flow induced by circular bubble plumes and water jets. *J Hydr Eng* 134(6):833–842
52. Lima Neto IE, Zhu DZ, Rajaratnam N (2008) Horizontal injection of gas-liquid mixtures in a water tank. *J Hydr Eng* 134(12):1722–1731
53. Lübcke HM, Rung Th, Thiele F (2003) Predication of the spreading mechanism of 3D turbulent wall jets with explicit Reynolds-stress closures. *Int J Heat Fluid Flow* 24:434–443
54. Margason RJ (1993) Fifty years of jet in cross flow research. Computational and experimental assessment of jets in cross flow. AGARD, Report CP-534
55. Mazurek KA, Christison K, Rajaratnam N (2002) Turbulent sand jet in water. *J Hydr Res* 40(4):527–530
56. Milgram H (1983) Mean flow in round bubble plumes. *J Fluid Mech* 133:345–376
57. Morchain J, Moranges C, Fonade C (2000) CFD modeling of a two-phase jet aerator under influence of a crossflow. *Water Res* 34(13):3460–3472
58. Morton BR, Taylor GI, Turner JS (1956) Turbulent gravitational convection from maintained and instantaneous sources. *Proc R Soc Lond A* 234:1–23
59. Motarjemi M, Jameson GJ (1978) Mass transfer from very small bubbles – the optimum bubble size for aeration. *Chem Eng Sci* 33(11):1415–1423
60. Mueller JA, Boyle WC, Pöpel HJ (2002) Aeration: principles and practice. CRC Press, New York
61. Newman BG, Patel RP, Savage SB, Tjio HK (1972) Three-dimensional wall jet originating from a circular orifice. *Aeronaut Quart* 23:188–200
62. Pande BBL, Rajaratnam N (1979) Turbulent jets in co-flowing streams. *ASCE J Eng Mech* 105:1025–1038
63. Pande BBL, Rajaratnam N (1977) An experimental study of bluff buoyant turbulent surface jets. *J Hydr Res* 15(3):261–275
64. Pani BS, Lee JHW, Lai ACH (2009) Application of Reichardt's hypothesis for multiple coflowing jets. *J Hydro-environ Res* 3(3):121–128
65. Papanicolaou PN, List EJ (1988) Investigations of round vertical turbulent buoyant jets. *J Fluid Mech* 195:341–391
66. Parr AD, Sayre WW (1979) Multiple jets in shallow flowing receiving waters. *ASCE J Hydraul Div* 105(HY11):1357–1374
67. Parthasarathy RN, Faeth GM (1987) Structure of particle laden turbulent water jets in still water. *Int J Multiph Flow* 13(5):699–716
68. Patel RP (1971) Turbulent jets and wall jets in uniform streaming flow. *Aeronaut Q* 22:311–326
69. Platten JL, Keffer JF (1971) Deflected Turbulent Jet Flows. *Trans ASME J Appl Mech* 38E(4):756–758
70. Pratte BD, Baines WD (1967) Profiles of the round turbulent jets in a cross flow. *J Hydr Div ASCE* 92:53–64
71. Pun KL, Davidson MJ, Wang HJ (2000) Merging jets in a co-flowing ambient fluid – a hybrid approach. *J Hydraul Res* 38(2):105–114
72. Rajaratnam N (1967) Plane turbulent wall jets on rough boundaries. *Water Power* 49:149–242
73. Rajaratnam N (1976) Turbulent jets. Elsevier Scientific, Amsterdam, 304 p
74. Rajaratnam N (1984) Turbulent jets and plumes, Lecture Notes, University of Alberta, Edmonton, Canada
75. Rajaratnam N (1985) An experimental study of bluff surface discharges with small Richardson number. *J Hydraul Res* 23(1):47–55
76. Rajaratnam N, Humphries JA (1984) Turbulent non-buoyant surface jets. *J Hydraul Res* 22(2):103–105
77. Rajaratnam N, Pani BS (1974) Three-dimensional turbulent wall jets. *Proc ASCE J Hydraul Div* 100:69–83
78. Rajaratnam N, Subramanya K (1967) Diffusion of rectangular wall jets in wide channels. *J Hydraul Res* 5:281–294
79. Rajaratnam N, Subramanyan S (1985) Plane turbulent buoyant surface jets and jumps. *J Hydraul Res* 23(2):131–146
80. Roberts PJW, Snyder WH, Baumgartner DJ (1989) Ocean outfalls I: Submerged waste field formation. *J Hydraul Eng* 115(HY1):1–25

81. Roberts PJW, Snyder WH, Baumgartner DJ (1989) Ocean outfalls II: Spatial evolution of submerged waste field. *J Hydraul Eng* 115(HY1):26–48
82. Rodi W, Spalding DB (1970) A two-parameter model of turbulence and its application to free jets. *Wärme Stoffübertragung* 3:85–95
83. Rostamy N, Bergstrom D, Sumner D (2010) An experimental study of a turbulent wall jet on smooth and rough surfaces. In: Nickels TB (ed) IUTAM symposium on the physics of wall-bounded turbulent flows on rough walls, Cambridge, UK
84. Rouse H, Yih CS, Humphreys HW (1952) Gravitational convection from a boundary source. *Tellus* 4:201–210
85. Rutherford JC (1994) River mixing. Wiley, West Sussex, 347 p
86. Schierholz EL, Gulliver JS, Wilhelms SC, Henneman HE (2006) Gas transfer from air diffusers. *Water Res* 40(5):1018–1026
87. Schlichting H, Gersten K (2000) Boundary layer theory. Springer, Berlin Heidelberg, 801 p
88. Sforza PM, Herbst G (1970) A study of three-dimensional incompressible turbulent wall jets. *Am Inst Aeronaut Astron J* 8:276–283
89. Sheen HJ, Jou BH, Lee YT (1994) Effect of particle size on a two-phase turbulent jet. *Exp Therm Fluid Sci* 8: 315–327
90. Singanetti SR (1966) Diffusion of sediment in submerged jet. *J Hydraul Div* 92(2):153–168
91. Smith SH, Mungal MG (1998) Mixing, structure and scaling of the jet in crossflow. *J Fluid Mech* 357:83–122
92. Socolofsky SA, Adams EE (2002) Multi-phase plumes in uniform and stratified crossflow. *J Hydraul Res* 40(6): 661–672
93. Tamburello DA, Amitay M (2008) Active manipulation of a particle laden jet. *Int J Multiph Flow* 34:829–851
94. Tang HS, Paik J, Sotiropoulos F, Khangaonkar T (2008) Three-dimensional numerical modeling of initial mixing of thermal discharges at real-life configurations. *J Hydraul Eng* 134(9):1210–1224
95. Tachie MF, Balachandrar R, Bergstrom DJ (2004) Roughness effects on turbulent plane wall jets in an open channel. *Exp Fluid* 37(2):281–292
96. Tian X, Roberts PJW, Daviero GJ (2004) Marine wastewater discharges from multiport diffuser. II: unstratified flowing water. *J Hydraul Eng* 130(12):1147–1155
97. Turner JS (1973) Buoyancy effects in fluids. Cambridge University Press, Cambridge, 367 p
98. Varley J (1995) Submerged gas-liquid jets: bubble size prediction. *Chem Eng Sci* 50(5):901–905
99. Verhoff A (1963) The two-dimensional turbulent wall jet with and without an external stream. Report 626, Princeton University
100. Wang HJ, Davidson MJ (2003) Jet interaction in a still ambient fluid. *J Hydraul Eng* 129(5):349–357
101. Whipple WJ, Yu SL (1970) Instream aerators for polluted rivers. *J Sanit Eng Div* 96(5):1153–1165
102. Wright SJ (1977) Mean behavior of buoyant jet in a crossflow. *J Hydraul Div ASCE* 103(HY5):499–513
103. Wright SJ (1984) Buoyant jets in density-stratified crossflow. *J Hydraul Eng ASCE* 110(5):643–656
104. Wüest A, Brooks NH, Imboden DM (1992) Bubble plume modeling for lake restoration. *Water Resour Res* 28(12):3235–3250
105. Wygnanski I, Fielder H (1969) Some measurements in the self-preserving jet. *J Fluid Mech* 38:577–612
106. Yan J, Luo K, Fan J, Tsuji Y, Cen K (2008) Direct numerical simulation of particle dispersion in a turbulent jet considering inter-particle collisions. *Int J Multiph Flow* 34: 723–733
107. Zhang W, Zhu DZ (2011) Near-field mixing downstream of a multiport diffuser in a shallow river. *J Environ Eng* 137(4): 230–240

Books and Reviews

- Cheremisinoff NP (ed) (1986) Encyclopedia of fluid mechanics, vol 2, Dynamics of single-fluid flow and mixing. Gulf, Houston, 1505 p
- Cheremisinoff NP (ed) (1986) Encyclopedia of fluid mechanics, vol 3, Gas-liquid flow. Gulf, Houston, 1535 p
- Cheremisinoff NP (ed) (1993) Encyclopedia of fluid mechanics, vol 2, Advances in multiphase flow. Gulf, Houston, 674 p
- Davies PA, Valente Neves MJ (eds) (1994) Recent research advances in the fluid mechanics of turbulent jets and plumes. Kluwer, Dordrecht, 514 p
- Fay JA (1973) Buoyant plumes and wakes. *Ann Rev Fluid Mech* 5:151–160
- Jirka GH (2004) Integral model for turbulent buoyant jets in unbounded stratified flow. Part 1: single round jet. *Environ Fluid Mech* 4:1–56
- Jirka GH (2006) Integral model for turbulent buoyant jets in unbounded stratified flow. Part 2: plane jet dynamics resulting from multiport diffuser jets. *Environ Fluid Mech* 6:43–100
- List EJ (1982) Turbulent jets and plumes. *Ann Rev Fluid Mech* 14:189–212
- Philip AA (1997) Jets, plumes and mixing at the coast. In: Allen PA (ed) Earth surface processes. Blackwell, Oxford, pp 241–266, Chapter 7
- Rodi W (1982) Turbulent buoyant jets and plumes. Pergamon, Oxford, 184 p
- Rubin H, Atkinson J (2001) Environmental fluid mechanics. Marcel Dekker, New York, 728 p
- Shen HH, Cheng AHD, Wang KH, Teng MH, Liu CCK (eds) (2002) Environmental fluid mechanics theories and applications. ASCE, Reston, 467 p
- Wood AW (2010) Turbulent plumes in nature. *Ann Rev Fluid Mech* 42:391–412
- Wood IR, Bell RG, Wilkinson DL (eds) (1993) Ocean disposal of wastewater. World Scientific, Singapore, 425 p

Tritium in the Environment

PHILIP A. DAVIS¹, DAN C. GALERIU²

¹AECL, Chalk River, ON, Canada

²Life and Environmental Physics Department,
“Horia Hulubei” Institute for Physics and Nuclear
Engineering, Bucharest-Magurele, Romania

Article Outline

Glossary

Definition of the Subject

Introduction

Properties and Species of Tritium

Behavior of Tritium in the Environment

Concentrations in Background Areas

Long-Term Average Concentrations Due to Routine
Releases

Time-Dependent Concentrations Due to Acute
Releases

Model Validation and Uncertainty

Future Directions

Appendix A

Bibliography

Glossary

Buried tritium Tritium in exchangeable positions in large biomolecules in the dry matter of environmental materials that is not removed by rinsing with tritium-free water

Deposition The process by which airborne contaminants are transferred from the atmosphere to the underlying surface (soil, plants, or water bodies)

Dose The amount of energy absorbed in the body, or in an organ or tissue of the body, due to exposure to ionizing radiation, multiplied by a weighting factor to account for the degree of biological damage inflicted by the type of radiation in question

Evapotranspiration The sum of evaporation (the movement of water to the air from sources such as the soil and water bodies) and plant transpiration (the movement of water within a plant and its subsequent loss as vapor to the air through stomata in the leaves)

Isotope One of two or more forms of an element differing from each other in atomic weight and in nuclear but not chemical properties

Isotopic effects Physical and chemical effects in plant and animal processes that arise due to differences in mass of different isotopes of the same element

Organically bound tritium (OBT) Carbon-bound and buried tritium formed in living systems through natural environmental or biological processes from HTO

Specific activity The amount of activity of a radioactive element per unit amount of the stable form of the element (Bq ³H/kg ¹H in the case of tritium)

Translocation The movement of a contaminant from one part of a plant or animal to another part

Tritium A radioactive isotope of hydrogen with one proton and two neutrons

Tritiated water (HTO) Water (or water vapor) in which a hydrogen atom is replaced by a tritium atom in some molecules

Definition of the Subject

Tritium is a radioactive isotope of hydrogen that can be released to the environment in small amounts during routine operation of nuclear facilities, and in higher amounts during some types of accidents. Tritium atoms exchange readily with hydrogen atoms in all environmental materials, including water. This means that tritium moves rapidly through the environment and enters the human body via inhalation of contaminated air and ingestion of contaminated food products. Once inside the body, tritium can be a constituent atom in any given water molecule and organic compound, imparting high local doses to cells that are radiosensitive. The study of environmental tritium is the study of the processes governing the behavior, transfer, and concentration of tritium in air, soil, plants, animals, and water bodies following releases to air and water. The understanding gained from such work is used to develop the ability to accurately estimate the dose to humans exposed to environmental tritium.

Introduction

Tritium (³H) is a radioactive isotope of hydrogen (¹H). It forms naturally in the upper atmosphere from the interaction of the neutrons in cosmic radiation with nitrogen and oxygen [1]. Very small amounts of natural tritium also arise from neutron capture by lithium in

rocks of the earth's crust. Anthropogenic sources of tritium include nuclear weapons testing, military production, fuel reprocessing plants, the manufacture of tritium lights, and medical and academic research that uses tritium as a radioactive label. Of more relevance in the present context, tritium is also produced during the operation of nuclear power reactors. All reactors control the fission process using neutron absorbers such as boron or lithium, which yield helium and tritium under neutron flux. Pressurized heavy water reactors have another route of tritium production, namely, the interaction of neutrons with the deuterium of the heavy water in the moderator and primary heat transport systems. Fusion reactors, in which tritium is the fuel, represent a future potential source of the isotope.

Nuclear reactors, either fission or fusion, are likely to play a major role in energy generation in a sustainable economy. This implies that small amounts of tritium will continue to be released routinely to air and water, and that larger releases are possible under accident conditions. Once in the environment, tritium is very mobile and can be effectively incorporated into biological systems, including the human body, where it can impart a significant dose to radiosensitive cells. The environmental behavior of tritium must be understood to estimate these doses, and to minimize risks to the environment and humans.

This entry reviews the current state of knowledge of environmental tritium. The properties of tritium and its chemical forms are discussed first, followed by a qualitative description of its behavior in terrestrial and aquatic systems. Quantitative sections on tritium concentrations in background areas and in the vicinity of routine and accidental releases are presented next from both observational and modeling points of view. The validity of environmental tritium models and the uncertainties in their predictions are discussed. This entry concludes with the identification of the main knowledge gaps and a discussion of future directions in environmental tritium research. Tritium dosimetry and human health effects are not covered here, but are addressed in the entry by Bundy et al. in this Encyclopedia.

Properties and Species of Tritium

The tritium nucleus contains one proton and two neutrons; the hydrogen nucleus also contains one proton

but no neutrons. Tritium decays to ^3He by emitting a beta particle (electron) and neutrino from one of its neutrons. The energy range of the beta particle is from 0 to 18.6 keV with an average energy of 5.69 keV. The radioactive half-life of tritium is 12.26 years.

Tritium readily exchanges with hydrogen and so is found in the multiplicity of environmental materials in which hydrogen occurs. The electrical configuration and chemical properties of tritium and hydrogen are similar but the two isotopes show small differences in some physical phenomena and some chemical reaction rates due to the difference in their atomic masses. Moreover, the presence of the β -decay energy associated with tritium provides the activation energy for some chemical reactions. Reactions involving tritium can occur at lower temperatures than those involving hydrogen.

Tritium is emitted to the environment from nuclear facilities in two main chemical forms: tritiated water (HTO) and tritiated hydrogen gas (HT). HTO dominates releases from fission reactors and from most existing tritium-handling facilities, and is the form most commonly found in the environment. HTO released to the atmosphere mixes with air moisture and exchanges with water pools in soil, plants, and animals. Exposure pathways to humans include inhalation, absorption through the skin, and food ingestion. HTO released with liquid effluents contaminates the receiving water body and is taken up by aquatic organisms [2]. Humans are exposed when they drink contaminated water and eat contaminated fish.

Plants and animals incorporate a fraction of their HTO content into organic compounds in isotopic-exchange or enzyme-catalyzed reactions [3, 4]. In exchange reactions, tritium bonds to oxygen, sulfur, phosphorus, or nitrogen atoms in the form of hydroxides, thiols, phosphides, and amines, respectively. Conventionally, this is termed exchangeable organically bound tritium (OBT). Exchangeable OBT is in equilibrium with the HTO in the plant or animal in question, and behaves as HTO in all respects. In enzyme-catalyzed reactions, tritium bonds to carbon chains of organic molecules. This is usually termed non-exchangeable OBT. This is the form of tritium that remains in dry biomatter that has been repeatedly washed with tritium-free water. Tritium so bound is more strongly attached than exchangeable tritium and

has longer retention times; such bonds are only dissolved during catabolic reactions.

Plant HTO is involved in the formation of OBT through various metabolic processes, primarily photosynthesis, that use plant water. OBT is formed in plant leaves but can be translocated to the edible parts of the plant and taken up by animals and humans through the food chain. Metabolic processes in aquatic plants also result in the formation of OBT, which is transferred to higher consumers in the ecosystem through ingestion. OBT taken into the body is broken down into more basic biological building blocks and then rebuilt into organic material in the body. About 50% of ingested OBT is converted to HTO during this process. In contrast, most ingested HTO remains as HTO in the body.

A small fraction of the tritium emitted by fission reactors is in the form of tritiated hydrogen gas. HT is not of radiological concern in itself since it is only weakly absorbed by the body and imparts a very low dose. However, HT released to the atmosphere diffuses into soil pore spaces, where it is oxidized to HTO by microorganisms containing the hydrogenase enzyme [5]. Some of the HTO so formed is taken up by plants and some is emitted to the atmosphere. Thus all of the exposure pathways discussed above for atmospheric HTO releases, including the formation and ingestion of OBT, are relevant for HT releases as well. The dose consequences of an atmospheric HT release are determined entirely by the HTO formed in the soil and its subsequent environmental behavior. Because HT is insoluble in water, there is no need to consider liquid HT releases.

Tritium can also be released to the environment in organic forms such as tritiated methane, tritiated pump oil, or tritiated pharmaceuticals. These tritiated organics can be inhaled and ingested, since they can be taken up to a limited extent by plants. They generally make a very small contribution to the total tritium dose to members of the public, but exceptions can occur for large releases [6].

Specific activity concepts are useful in understanding environmental tritium behavior. Specific activity (SA) is defined as the radioisotope activity per unit mass of the stable element ($\text{Bq } ^3\text{H}/\text{kg } ^1\text{H}$ in the case of tritium). Because tritium atoms exchange so rapidly with hydrogen in aqueous parts of the environment (including the water of living organisms), the behavior

of tritium mimics that of hydrogen in physical and biological processes. In particular, the flux of tritium from one environmental compartment to another is determined by the flux of hydrogen. If two compartments with different specific activities are allowed to come into equilibrium, the resulting SA of both compartments (which is determined by the total tritium and total hydrogen contents of the compartments) will be equal, with a value that lies between the two initial SAs. This implies that there can be no bioaccumulation of HTO in the environment. Tritium differs from most other radionuclides in this respect. For example, the cesium concentration in fish can exceed the cesium concentration in the water in which the fish swim, but this is not the case for HTO.

In theory, the SA model applies to OBT as well as to HTO, and it does in practice when long-term average concentrations are considered in a system where OBT is formed metabolically from HTO. However, the long biological half-life of OBT relative to that of HTO occasionally results in the apparent violation of SA concepts. The OBT concentration in a plant exposed periodically to an airborne HTO plume will be higher than the atmospheric HTO concentration when the plume is not present. Similarly, OBT concentrations will exceed HTO concentrations in organisms if the OBT appears through direct incorporation of tritiated organic material with high SA. For example, the OBT/HTO ratios measured in fish from Cardiff Bay in the UK are high because the fish ingest tritiated organics released to the water column from a nearby pharmaceutical plant [6]. These are not examples of the bioaccumulation of tritium, but rather the manifestation of the mismatch in the biological half-lives of HTO and OBT. These ideas are discussed elaborately in [section "Nonequilibrium Conditions."](#)

Tritium specific activity can be formulated in terms of the tritium concentration in water (in Bq L^{-1}) rather than the ratio of tritium activity to the mass of hydrogen in a given compartment. There is a one-to-one correspondence between these two quantities and it is the concentration in water that is normally measured. In the case of OBT, the concentration in Bq L^{-1} is determined as the activity in the water equivalent of the dry matter (the water produced by complete combustion of the dry material).

The low energy of the tritium beta particle means that the radiation cannot penetrate human skin. Thus, tritium does not impart an external dose from deposited radioactivity or immersion in contaminated air. Tritium can cause damage to organisms only if taken into the body through inhalation, ingestion, or absorption through the skin. Once inside the body, tritium can be a constituent atom in any given water molecule and organic compound. HTO and exchangeable OBT are retained in the body with a biological half-life of approximately 10 days. HT can dissolve in the blood but is exhaled within a few minutes. Doses from the inhalation of HT are therefore about 25,000 times lower than doses from inhaled HTO. On the other hand, non-exchangeable OBT has a biological half-life of about 40 days and imparts a dose that is more than twice that of HTO. Since the tritium atoms in non-exchangeable OBT are fixed in place, the subcellular location of a given atom is of importance in determining the effects of its beta decay, and high absorbed doses can occur locally in cells that are radiosensitive [7]. Furthermore, because of its low initial energy and short range, the average density of ionization produced by the passage of a tritium beta particle is higher than that for a moderate/high energy particle.

More information on tritium dosimetry and the dose consequences of tritium releases to the environment can be found in the entry by Bundy et al. in this Encyclopedia.

Behavior of Tritium in the Environment

Hydrogen is ubiquitous in the environment and is part of many chemical compounds, including water and most organic materials. As an isotope of hydrogen, tritium enters freely into these compounds and its movement through the environment can be inferred from the cycling of hydrogen. As a result, tritium is unique among radionuclides in a number of respects:

- In aqueous systems, tritium moves as a nonreactive, non-absorbed constituent with the bulk water flow. Accordingly, the environmental transport of tritium is governed in large part by local and global hydrologic cycles.
- As a gas, tritium moves in response to its own vapor pressure gradient and can, under some circumstances, move against the water vapor flux.

- Tritium deposited from the atmosphere to soil and plants is readily recycled back to the atmosphere via evapotranspiration, forming a secondary airborne plume.
- The processes responsible for tritium transfer have timescales as short as minutes. Tritium can be rapidly taken up by organisms but just as rapidly lost. As a result, tritium transfer is highly dependent upon the environmental conditions prevailing at the time and place of release and the time of measurement.
- Tritium can be effectively incorporated into biological systems, including the human body, as organically bound tritium.
- Although tritium is three times heavier than hydrogen, it usually occurs as part of larger molecules. Therefore, isotopic effects, although present, are not important in environmental tritium transport, except in OBT formation.
- Tritium is transferred through the environment and through food chains without bioaccumulation in any compartment.

These unique features must be taken into account in understanding the environmental transport of tritium and in estimating the radiological consequences of releases to the environment.

Cycling of HTO Among Air, Soil, and Vegetation

On release to the atmosphere, HTO is advected by the mean wind and diffused by turbulent eddies in the same way as other radionuclides. A receptor close to a tritium source does not experience steady air concentrations even if the release rate is constant and continuous. Changes in wind speed, stability, and particularly wind direction induce changes in air concentrations on hourly timescales.

HTO can be lost from the atmosphere by wet or dry deposition to soil, plants, or surface water. As a vapor, HTO can diffuse directly into the soil pore space and into the leaf stomates [8, 9]. This deposition rate is driven by the gradient between the concentration of HTO in the atmosphere and in the air within the pore space or within the sub-stomatal cavities. HTO deposition can occur even when the water vapor flux is upward, as long as the HTO partial pressure gradient is downward. This happens, for example, when an

airborne plume passes over a relatively uncontaminated surface on a clear summer day when evapotranspiration is in progress [10]. The deposition rate depends on the diffusion resistances [11] that the tritium molecules encounter in passing from the air to the underlying surface. These include the aerodynamic resistance of the atmosphere, the boundary layer resistance of the laminar sub-layer close to the plant or soil surface, and the stomatal and surface resistances of the plant and soil matrix, respectively.

HTO is also lost from the air to underlying surfaces by wet deposition. HTO molecules diffuse through the air and through the surface of individual raindrops or snowflakes and are carried to the ground with the precipitation. The amount of washout depends on the diffusion resistances and on the concentration of HTO in air moisture (C_{am}). The HTO concentration in precipitation is usually less than C_{am} because the precipitation, which is initially uncontaminated as it starts its fall, does not come into equilibrium with air moisture in the small amount of time needed to pass through the plume.

When the wind blows the plume away from a receptor that previously experienced a period of deposition, the HTO partial pressure gradient and the diffusive flux will immediately reverse direction and HTO will be rapidly lost from soil and plants to the atmosphere by evapotranspiration. The rate of loss is determined by the HTO gradient and the aerodynamic, boundary layer, stomatal, and surface resistances discussed above in connection with deposition. In these circumstances, the soil and plants represent a ground-level area source that generates a secondary airborne HTO plume.

The rate of HTO reemission from soils depends strongly on the HTO distribution with depth that resulted from earlier deposition. The distribution determines the initial availability of HTO at the soil surface and the subsequent resupply from deeper layers by diffusion and advection. The tritium profile in soil following short-term dry deposition of HTO takes the form of an error function with a scaling length of less than 1 cm [12]. Most of the tritium remains near the soil surface and reemission occurs quickly once the airborne plume has passed. In contrast, the profile resulting from wet deposition tends to be fairly uniform with depth since rain and any tritium it contains

is able to penetrate to lower levels of the soil. In this case, reemission occurs slowly because relatively small amounts of HTO are found near the surface and because the high soil water content retards upward HTO diffusion from lower layers.

Because tritium deposited under dry conditions tends to be quickly reemitted when the plume moves off a given location, the persistent HTO inventory in soil arises largely through wet deposition. Tritium not returned to the atmosphere by evaporation moves through the soil by vapor transport or by the mass flow of liquid water. Vapor transport occurs by diffusion through the air-filled pores at a rate determined by the HTO concentration gradient and is usually important only in dry soils. The mass flow mechanism is common to other radionuclides but is more efficient for tritium, which undergoes no retardation and moves at essentially the same rate as water under the influence of gravity and capillary forces. Tritium is lost from the soil by drainage and root uptake, the latter mechanism depending on the rooting depth of the vegetation. The HTO concentration in soil water depends on all of the processes discussed above as well as on the soil water content. The soil concentration tends to be lower than C_{am} because precipitation, which is the primary source of soil HTO, has a concentration lower than C_{am} .

Over long time periods in humid climates, tritiated water moves gradually downward through the soil profile as it is displaced by subsequent precipitation. It eventually passes below the root zone where it may enter regional groundwater systems. It is then lost to the local biosphere unless the aquifer discharges to a surface water body or is tapped by a well.

Tritium is carried from the soil into plant roots with the bulk water flow driven by transpiration. The tritium concentration in the transpiration stream is a weighted average of the concentrations in the soil layers from which the water is drawn. The HTO concentration in leaves is determined by the concentration in the transpiration stream, by the uptake or loss of tritium by vapor diffusion through the stomata, and by the volume of water in the plant. Under steady conditions, the concentration in the plant lies between that of the soil and air with a magnitude that depends on the atmospheric humidity and the air/leaf temperature difference [13]. Concentrations in fruit tend to be lower than concentrations in foliage due to a slower

water turnover rate in fruit. Concentrations in root crops are also low since these crops draw a relatively greater fraction of their water from the soil than they do from the air.

Plant concentrations do not remain static during the night even though little transpiration occurs and the diffusion resistance is high because the stomata are closed. Experiments suggest that tritium continues to be taken up by the plant leaves during the night, although at a lower rate than during the day [14, 15].

HTO concentrations in air, plants, and soil at a given location near a tritium source are in a highly dynamic state [16]. When the atmospheric plume is present, air concentrations are high, plant concentrations build up rapidly and then level off, and surface soil concentrations increase steadily. When the atmospheric plume moves off, air concentrations decrease abruptly but remain measurable because of reemission of deposited tritium. Similarly, plant concentrations drop off rapidly to levels that reflect concentrations in the transpiration stream. Surface soil concentrations also decrease quickly, although concentrations at deeper levels change only slowly with time. The cycle repeats when the wind brings the plume back over the location in question. In general, concentrations in any one part of the system are not in equilibrium with concentrations in any other part at any time, as they should according to specific activity concepts. This can create problems in interpreting tritium field measurements, particularly when sampling in the various compartments is done at different times and/or over different averaging periods. For example, the HTO concentration in vegetation can be much lower than the air concentration (if the plant sample is taken just after the airborne plume reaches the site) or much higher (if the sample is taken just after the plume moves off the site). These are not examples of isotopic effects or tritium bioaccumulation, but simply artifacts of measurements made in a transient system.

Organically Bound Tritium Formation in Plants

Photosynthesis is the most important process of tritium incorporation into organic compounds during daylight hours. Non-exchangeable OBT formed by photosynthesis in green plants appears initially in carbohydrates [3]. Subsequent polymerization results in

the incorporation of OBT in such complex molecules as polysaccharides (starch and cellulose), proteins, lipids, and nucleic acids. The amount of OBT produced depends on a large number of environmental factors and plant parameters including light levels, oxygen and carbon dioxide concentrations, temperature, air circulation, and water supply, all of which show considerable diurnal and seasonal variations. OBT concentrations are reduced by slow conversion back to HTO by maintenance respiration and, following an acute exposure, by subsequent plant growth. NCRP [17] reports a biological half-life of non-exchangeable OBT in plants of about 25 days, but this value includes the effects of growth dilution. OBT makes up a few percent of the total tritium activity in most plants but up to 90% in grains, which have a high organic content.

OBT can also be produced in the dark, although at rates 3–10 times lower than those resulting from photosynthesis [14, 15, 18, 19]. The processes involve the metabolic turnover and synthesis of organic compounds such as proteins, oils, alkaloids, and vitamins using the energy of respiration. Plant tissue is dynamic, breaking down, and rebuilding continuously, and tritium atoms from HTO can be incorporated into non-exchangeable sites during the rebuilding phase.

OBT is formed only in the green parts of plants, but can be translocated to roots, stems, and edible fruits. The compounds formed by photosynthesis at the beginning of the vegetative period (including those containing OBT) are used primarily for growth of leaves, stems, and roots. Later in the growing season, usually after flowering, most of the assimilates are used in the development of fruits and tubers. The amount of OBT appearing in edible plant parts is therefore greatest if the exposure to atmospheric tritium occurs after flowering [20]. Diabate and Strack [21] observed that the major factor influencing translocation is the growth stage of the plant, while light conditions do not play a significant role.

The large difference in mass between hydrogen and tritium gives rise to significant isotopic effects in OBT formation. Tritium is discriminated against in the cleavage of the O–H bond of tritiated water during photosynthesis, so that less tritium than hydrogen is incorporated into organic molecules. In contrast, once fixed to carbon, tritium is cleft more slowly than C–H bonds. The net result is an OBT concentration that is

20–40% lower than the HTO concentration in the water to which the plant is exposed [22, 23].

Recently, Baumgärtner, and Donhärsl [24] and Baumgärtner [25] identified another form of OBT, called buried tritium, which is defined as the tritium in exchangeable positions in large biomolecules in dry matter that is not removed by rinsing with tritium-free water. Such tritium is not carbon bound but is simply folded into large molecules that are not accessible for exchange with tritium-free water. Buried tritium contributes to the concentration in the traditional experimental determination of OBT but it quickly exchanges with hydrogen atoms in the body and acts as HTO rather than OBT following ingestion. The rapid increase in plant OBT concentrations observed in the first few hours following acute HTO exposure in some experiments may be evidence of buried tritium forming by exchange through non-metabolic processes [15].

The nature of OBT was debated by the Tritium and Carbon-14 Working Group of the International Atomic Energy Agency's EMRAS (Environmental Modeling for Radiation Safety) program [26], which settled on the following definition: "OBT is carbon-bound and buried tritium formed in living systems through natural environmental or biological processes from HTO. Other types of organic tritium (e.g. tritiated methane, tritiated pump oil, radiochemicals and so on) should be called tritiated organics, which can exist in any chemical or physical form." This definition was clarified through a number of notes, and the full definition is given in [Appendix A](#).

The EMRAS definition is consistent with the quantity measured in the traditional analytical determination of OBT, namely the tritium concentration in the combustion water produced by the oxidation of dry matter that has been washed with tritium-free water [27]. The definition was deliberately chosen to maintain this connection in order to preserve the meaning of past analyses and current analytical techniques. The definition does not include non-exchangeable OBT but does include buried tritium. It also distinguishes between organically bound tritium that is formed in the environment and tritiated organics, which are produced through industrial processes.

The concept and significance of buried tritium is still under debate. Baumgärtner and Donhärsl [24] and Baumgärtner [25] suggest that buried tritium makes up

50% or more of what is traditionally measured as OBT. Lower fractions, typically 5% and at most 20%, have been reported elsewhere [26]. In the face of this discrepancy, the question remains open pending new experimental data. If buried tritium is a reality and behaves as non-exchangeable OBT in the body, OBT doses may be overestimated by present OBT measurements.

Uptake by Terrestrial Animals

Tritium is taken into the bodies of animals (including human beings) through inhalation and ingestion of food and water [28, 29]. Uptake also occurs by absorption through the skin, which is a minor pathway for animals with their thick skins. Losses occur through respiration, sweating, excretion, and (in the case of cows) milking. Tritium concentrations in the body depend on rates of uptake and loss, and on concentrations in the air, drinking water, and food taken in by the animal. These concentrations may vary widely since animals are mobile and may draw their air, food, and water from different parts of the environment.

HTO diffuses freely and rapidly across all cell membranes and equilibrates with body fluids within minutes. When exposure stops, the HTO concentration in the body decreases with a biological half-life of 3–10 days for the vast majority of the tritium, which behaves as HTO in the body water [30], and a half-life of 40 days for the small amount incorporated as OBT into body dry matter [31]. These timescales suggest that tritium concentrations in animals fluctuate over time but do not follow day-to-day changes of concentrations in their environment.

Non-exchangeable OBT in the body arises principally from assimilation of ingested OBT in foods [32]. As much as 50% of the OBT in food fed to animals appears as OBT in the animal itself, attached to biochemical compounds such as amino acids, sugars, protein, starches, lipids, and cell structural materials. A small fraction (<5%) of HTO taken into the body is irreversibly incorporated into organic molecules through the enzymatic reactions involved in carbohydrate metabolism [33]. Twenty times more OBT is incorporated into milk fat and casein after OBT ingestion than after HTO intake [30, 34]. The extent of incorporation depends on the length of exposure and

the metabolic activity, which differs from organ to organ. Thus, OBT can be localized in the body in a relatively small number of cells and at relatively high concentrations. Fats and proteins retain more of their OBT during digestion than do carbohydrates. However, most OBT ingested with food is metabolically oxidized to HTO during digestion and assimilation and enters the compartment of free body water.

OBT has a longer retention time than tritiated water in animals. When exposure stops, OBT is lost slowly from the body as a result of metabolic oxidation to HTO, which is then excreted. The biological half-life of this process is variable. van den Hoek et al. [34] measured values ranging from 60 to 330 days, whereas Galeriu et al. [35] suggest values of 88 days for cows, 75 days for calves, >100 days for pigs, and >190 days for goats. Hill and Johnson [36] reviewed OBT retention half-lives for humans and found values ranging from 20 to 226 days. Based on studies in which adult rats were fed dried flounder containing OBT, Hodgson et al. [31] proposed that 70% of the tritium in humans would be excreted with a 10-day half time and 30% with a 100-day half time. The biological half-life varies with the type of molecule (carbohydrate, protein, or lipid) since different molecules have significantly different metabolic turnover rates. The half-lives for animals reflect protein and lipid metabolism and are longer than half-lives for plants, which are composed primarily of carbohydrates [37, 38]. The constant synthesis of organic compounds in dairy cows during milk production results in the continuous formation of OBT, which is excreted in milk, urine, and feces at a rate only slightly slower than that of HTO.

Conversion of HT to HTO

Small amounts of tritiated hydrogen are released from heavy water reactors and larger amounts from tritium recovery facilities. HT is biologically inert, and direct exposure results in very low radiological doses. However, HT can diffuse into the soil and be converted to HTO by enzyme-mediated reactions. The converted HT is subsequently transported as HTO, which dominates doses from an HT release. Therefore, deposition and conversion are the only HT-specific processes that need to be explicitly considered in assessing the consequences of an HT release.

The oxidation of HT in the soil is mediated by hydrogenase, an enzyme that exists freely or in association with many bacteria of various taxonomic groups, in many eucaryotic algae, in some protozoan microorganisms, and in several mosses [39, 40]. Hydrogenase is able to break the bonds in the HT molecule and then oxidize it to tritiated water. Oxidation rates are independent of soil particle composition, organic content, pH, or chemical properties [41]. The number of HT-oxidizing organisms decreases with depth in the soil [5] but is sufficiently great to provide biological reaction sites for all HT molecules at all depths. Thus the rate of deposition is usually not limited by the number or activity of conversion sites, since the soil has the capacity to convert to HTO essentially all the HT that diffuses into it. Rather, the rate is governed by processes that control access of HT to the reaction sites on pore walls. These processes include diffusion both into the soil and within the pore air, both of which are affected by soil water content, which is the most influential variable determining deposition rate. The water content controls the available air-filled pore space and thus the rate of HT diffusion within the soil. Water content also affects the degree to which soil particles are covered by water films, which restrict access of the relatively insoluble HT to reaction sites located on soil particle surfaces. The HT deposition rate therefore decreases with increasing soil water content.

HT deposition rates also decrease rapidly as soil moisture drops below 3% [42]. Biological activity is diminished in extremely dry soils, and in this case, HT deposition is limited by the reduced number of conversion sites. In temperate regions, however, surface soil water contents typically exceed 4% and oxidation capacity is not an issue except during prolonged dry spells. Oxidation activity and HT deposition rate also decrease as soil temperatures drop below freezing [43], although some seasonal variability may be due to changes in soil moisture content rather than to variations in temperature. Organic content affects available pore space and pore size distribution and thus can influence diffusion and HT deposition rate [5]. Soils with vegetation cover show higher deposition rates than bare soils, suggesting that root systems enhance soil porosity (and hence diffusion) and increase microbial activity.

The combined effects of diffusion and microbial oxidation result in an HTO profile that decreases

exponentially with depth in the soil [44]. The depth at which the HTO concentration drops to $1/e$ of its value at the surface (the scaling length) is typically between about 1.5 and 3.0 cm, which implies that most of the HT is oxidized within the top 6 cm of the soil. The diffusion rate of HT in soil is faster than that of HTO, resulting in a lower surface concentration and a deeper profile. The initial reemission rate for HT deposition is therefore about three times lower than it is in the case of an HTO source. For an HT release, the secondary airborne HTO plume appears immediately, even in the presence of the initial HT cloud, because the HTO concentration gradient is always directed upward.

The HT deposition rate is usually expressed quantitatively as a deposition velocity (v_d), defined as the HT deposition flux per unit horizontal area of soil divided by the concentration of HT in air. Numerous experiments have demonstrated that the deposition velocity of HT to soil is usually in the range 10^{-4} to 10^{-3} m s $^{-1}$ for soils of typical water content under nonfreezing conditions. Deposition velocities decrease with increasing water content and colder temperatures, dropping to about 10^{-5} m s $^{-1}$ for very wet or frozen soils. Good agreement exists between data obtained using chambers in the laboratory [45, 46], chambers in the field [43, 44, 47], and measurements made in the open environment [48–50]. Local inhomogeneities in soil conditions and vegetation cover can result in a tenfold variation in v_d within a natural field.

HT can be taken up and converted to HTO by foliage [40, 51], and gas-phase oxidation can occur in the atmosphere [52], but the amount of HTO produced by either of these processes is insignificant compared to the amount generated in the soil. Biological conversion of HT may proceed relatively rapidly within at least some types of soil litter, with deposition velocities in the lower range of those to soil [53, 54].

Aquatic Processes

Tritium released to a surface water body is transported by the mass flow of the receiving waters and its concentration is determined by dilution, dispersion, and evaporation. Of these processes, the only one not experienced by most other radionuclides is evaporation, but this can have a large influence on HTO concentrations in some systems [55]. Because fresh water systems vary

widely in their characteristics, concentrations depend strongly on the properties of the specific site. An aquatic receptor close to a tritium source may not experience steady water concentrations even if the release rate is constant and continuous. Changes in stratification, current speed, and (in lakes and oceans) current direction may cause water concentrations to fluctuate on daily timescales. Moreover, fish may experience changing water concentrations as they pass into and out of the tritium plume during their movements through the water body.

Most aquatic organisms are totally immersed in water and have high water exchange rates. Uptake of HTO is therefore very quick, and concentrations in tissue become equal to water concentrations within minutes or hours [56, 57].

The process of OBT formation is much the same in aquatic plants as it is in terrestrial plants. Exchangeable and non-exchangeable OBT are formed through photosynthesis during the day and through a variety of metabolic processes in the dark. The OBT can be translocated from its place of formation to storage organs in the plant, and is degraded to HTO through respiration. Isotopic effects also play a role in determining OBT concentrations. A small fraction of the HTO taken in by aquatic animals is converted to OBT in anabolic processes. Metabolic processes in the animal depend on water temperature, and seasonal effects should be taken into account in estimating OBT formation [58]. Fish metabolism slows down appreciably when the fish growth rate becomes less than 10% of the optimum growth rate, which occurs for water temperatures below 6–8°C. Elimination of OBT from fish tissues is relatively rapid for small fish, with biological half-lives in the range of 5–25 days [56, 59]. Data for larger fish are missing but longer values are expected.

Tritium reaches the oceans via direct release, via discharge from rivers and groundwaters, and via deposition from the atmosphere. It quickly spreads through the surface mixed layer, which is about 100-m thick, and is reduced to very low concentrations by the huge water volumes available for dilution [60]. Exchange with deeper layers is slow and any tritium transported downward joins oceanic circulation and decays away before the water reappears on the surface. Little is lost to the atmosphere since the low water concentrations mean that the diffusion gradient is usually directed

downward and deposition is favored over evaporation. Marine organisms take up tritium in the same way that fresh water organisms do, but concentrations are extremely low.

Environmental Tritium Concentrations

All of the environmental transport processes discussed above operate regardless of whether the tritium source is short-term or continuous, local or global. However, the resulting environmental concentrations depend very much on these factors. The following three sections discuss environmental tritium concentrations in background areas, near continuous sources, and near time-dependent sources, respectively.

Concentrations in Background Areas

In the present context, a background area is an area remote from local tritium-handling facilities. The only significant tritium sources in such areas are those that operate on global scales. These include cosmogenic production, neutron capture by lithium in rocks, and the residue of fallout from atmospheric nuclear weapons testing in the late 1950s and early 1960s. All of these sources together result in global background tritium concentrations of about $2\text{--}3 \text{ Bq L}^{-1}$ in air moisture, precipitation, surface waters, and the aqueous phase of living organisms [1]. This is an extremely low abundance, corresponding to roughly 1 tritium atom in 10^{17} hydrogen atoms. Because of the large spatial scale of sources in background areas, concentrations tend to be constant in time and space and are independent of the averaging time used to measure them.

Long-Term Average Concentrations Due to Routine Releases

Nuclear power reactors, fuel reprocessing plants, and tritium light factories all routinely release small amounts of tritium on a continuous, or quasi-continuous, basis. A receptor close to such a source does not experience steady air concentrations even if the release rate is constant. Changes in meteorological conditions, particularly wind direction, induce changes in air concentrations on hourly timescales. However, it is not instantaneous concentrations that are of interest for

routine releases since the doses they impart are very low. Rather it is the long-term average concentrations, which are used to ensure that annual dose limits are not exceeded, and which are generally steady from year to year. Thus, measurement programs and models for predicting environmental concentrations around continuous sources are generally designed to generate annual average values. Concentrations can be 10–100 times higher than background levels at the exclusion zone of tritium-handling facilities, and higher still closer to specific sources. Concentrations due to atmospheric releases of HTO, atmospheric releases of HT, and aquatic releases of HTO are discussed in turn below.

Atmospheric Releases of HTO

Observations Long-term average concentrations in the various environmental compartments at a given location near a continuous atmospheric tritium source are in equilibrium, but are not equal. Concentrations in precipitation are lower than those in air moisture because precipitation, which is initially uncontaminated, usually does not come into equilibrium with air moisture in the small amount of time needed to fall through the airborne plume. Concentrations in soil are similar to those in precipitation since the persistent HTO inventory in soil arises largely through wet deposition. Observations of long-term average concentrations in air moisture (C_{am}) and soil water (C_{sw}) suggest that values of the ratio $C_{\text{sw}}/C_{\text{am}}$ lie in the range 0.1–0.4 [61–63]. The value depends upon a number of local factors, notably the frequency with which rain falls when the airborne plume is present at the site of interest.

Plants take up tritium from both air moisture and soil water. The relative contribution of these compartments to the tritium concentration in the plant depends on the relative humidity (RH) of the atmosphere [13, 64]. When $\text{RH} = 1$, the transpiration stream shuts down and no soil tritium is carried to the leaves. In this case, the tritium in the leaves is due entirely to transfer from the air via diffusion through the stomates, and the leaf concentration comes into equilibrium with the concentration in air moisture. As the humidity drops, transpiration brings low concentration water up to the leaves, diluting the input from

the air. In the limit $RH = 0$ (a physically impossible situation), there is no HTO in the air and no transfer from air to leaf, and the plant concentration equals the concentration in the transpiration stream. The long-term average value of RH in temperate climates is about 75% and the plant HTO concentration lies closer to the concentration in air moisture than to that in soil water.

The relative contribution of air moisture and soil water to the tritium concentration in the plant also depends on the part of the plant under consideration. The partitioning in terms of relative humidity applies specifically to plant leaves that, under normal circumstances, draw the majority of their tritium from the air. The balance changes for fruit, tubers, or root crops, which draw a larger fraction of their tritium from the soil [65]. The concentration in tubers and root crops is close to that in soil water. The concentration in fruit water is intermediate between the concentration in leaves and soil water.

Observations of long-term average tritium concentrations in air moisture and plant water (C_{pw}) show that values of the ratio C_{pw}/C_{am} exhibit considerable scatter, with a mean of 0.68 for leafy and non-leafy vegetables [66]. Lower values apply to root crops and fruit, which draw more of their water from the soil.

Since most plant dry matter is formed in the presence of water, the plant OBT concentration is similar to the HTO concentration. However, due to the larger mass of tritium compared to hydrogen, isotopic fractionation results in an OBT specific activity that is lower than that for HTO [22]. The presence of exchangeable hydrogen in the plant combustion water also contributes to the relatively low dry weight concentration.

Because HTO and OBT have very different formation and clearance times in plants (a few hours and several weeks, respectively), and because tritium concentrations in air vary rapidly in time due to fluctuations in meteorological conditions, it is difficult to obtain reliable estimates of the long-term OBT/HTO ratio from field data. The most dependable values come from controlled laboratory experiments, where the plant is exposed to an HTO concentration that is held constant or monitored continuously. Table 1 summarizes literature values of OBT/HTO ratios obtained under controlled conditions. Although the number of

Tritium in the Environment. Table 1 Empirical values of the OBT/HTO ratio in plants obtained under controlled laboratory conditions

Plant type	OBT/HTO ratio	Reference
Maize	$0.51^a \pm 0.03$ 0.66 ± 0.01	[23] [67]
Barley	$0.42^a \pm 0.01$ 0.55 ± 0.01	[23] [67]
Alfalfa	$0.58^a \pm 0.02$	[22]
Mean	0.54 ± 0.09	

^aThe concentrations measured in these experiments as "OBT" consisted of both exchangeable and non-exchangeable organically bound tritium. To eliminate the contribution of exchangeable OBT, the OBT/HTO ratios reported in the publications were corrected by subtracting 0.22, which represents the proportion of exchangeable hydrogen in the total hydrogen of the samples [68].

data points is small, all values are less than 1, with a mean of 0.54 for the crops considered (maize, barley, and alfalfa).

Concentrations in animal products reflect the concentrations in the drinking water and various foods that make up their diet. Animal concentrations are generally lower than concentrations in air or plants for two reasons. First, some food products may be imported from sites remote from the source and be uncontaminated. Second, the tritium inventory is diluted by drinking water, which is usually drawn from a large water body with mainly uncontaminated input.

Models Most models designed to calculate long-term average tritium concentrations avoid the explicit simulation of the individual transport processes and short-term dynamics that are characteristic of environmental tritium behavior. To take these factors into account would require large amounts of computing time and site-specific data, and would result in large uncertainties in the predicted concentrations. Instead, it is customary to use simpler models that describe the integrated response of the system, and for which appropriate generic data exist. These models have traditionally been based on the specific activity approach, in which tritium is assumed to mix with hydrogen in a source compartment, resulting in a certain SA.

Any organism drawing hydrogen from this compartment draws tritium in proportion, and attains the same SA as the source compartment. If the organism draws hydrogen (and tritium) from more than one compartment, the SA in the organism is a weighted average of the specific activities in these compartments.

The general mathematical expression of this approach is given in Eq. 1. The specific activity SA_j in a compartment j that draws an element S from N other compartments is given by:

$$SA_j = \frac{A_j}{S_j} = \frac{\sum_{i=1}^N f_i \cdot D_{Fi} \cdot A_i}{\sum_{i=1}^N f_i \cdot S_i} \quad (1)$$

In Eq. 1, A_i is the concentration of the active form of the element in compartment i (Bq kg^{-1}), S_i is the concentration of the stable form (g kg^{-1}), and f_i is the flux of S from compartment i to compartment j (g d^{-1}). D_F is an isotopic discrimination factor that is introduced to allow for the possibility that the stable and active forms may have significantly different masses and therefore different transfer rates between compartments. The concentration of the radioisotope in compartment j is found by multiplying SA_j by the concentration of the stable element in the compartment.

For an atmospheric release of HTO, it is the concentration in air moisture C_{am} (Bq L^{-1}) that drives the transfer of tritium in the environment. The tritium concentration in air (C_a , Bq m^{-3} air) is related to C_{am} through the atmospheric absolute humidity q (kg m^{-3}):

$$C_a = q C_{am} \quad (2)$$

The value of C_a is generally known, either through measurement or through calculations involving an atmospheric dispersion model such as the sector-averaged Gaussian plume model. Care should be taken in choosing the air concentration used to drive the model. Plants take up tritium through their leaves primarily during daylight hours in the growing season, and air concentrations used to model this process should be based on meteorological data for this period. On the other hand, concentrations in soil and surface water should reflect annual average air concentrations,

since deposition occurs throughout the day and throughout the year.

In its simplest form, the SA model assumes that the tritium concentration in all environmental compartments (including man) is equal to the tritium concentration in air moisture at the location of interest [69]. However, as discussed above, the tritium concentration in the environment is usually less than that in air, as isotopic exchange with relatively uncontaminated water pools results in progressive dilution of the tritium in compartments farther along the exposure pathways. The net effect of these dilutions is a dose that is a factor two to three lower than that predicted on the assumption of full specific activity equilibrium in all compartments.

There are a number of ways to obtain more realistic estimates of long-term average concentrations in the environment. In the simplest of these, concentrations in soil and plants can be deduced directly from the concentration in air moisture using observed values of the ratios C_{sw}/C_{am} and C_{pw}/C_{am} . Similarly, the OBT concentration in plants can be obtained from the plant HTO concentration and the observed OBT/HTO ratio. This approach will give reasonably good estimates at sites where meteorological conditions, soil properties and plant species are similar to those at the sites where the ratios were observed, but may not perform as well otherwise.

A number of more detailed models have been proposed for calculating steady-state concentrations in the various environmental compartments. In a common approach to estimating soil water concentrations, the amount of HTO deposited (ω_w , Bq m^{-2}) in the course of one rainfall event that lasts Δt s is first calculated from

$$\omega_w = \Lambda Q \exp(-\Lambda x/u) \Delta t / (x u \Delta \theta), \quad (3)$$

where Λ is the washout coefficient (s^{-1}), Q is the release rate (Bq s^{-1}), x is the downwind distance from source to receptor (m), u is the wind speed (m s^{-1}), and $\Delta \theta$ is the angular width of the sector (radians), assuming that a sector-averaged dispersion model has been used to calculate air concentrations. The total amount of activity deposited per year is found by summing the amounts deposited during each rainfall occurring in that year. Soil concentrations are calculated by dividing ω_w by the annual precipitation. Only HTO that is

deposited with precipitation is assumed to be retained in the soil since dry-deposited HTO is quickly reemitted to the atmosphere by evaporation.

Belot [61] proposed a more complex model for soil water concentrations that accounts for the contributions of both wet and dry deposition:

$$C_{sw} = \frac{v_d C_a + F_w}{v_e \rho_s + I_r} \quad (4)$$

Here v_d is the exchange velocity of HTO from air to soil (m s^{-1}), F_w is the average flux density of tritium wet deposition ($\text{Bq m}^{-2} \text{s}^{-1}$), v_e is the exchange velocity from soil to air (m s^{-1}), ρ_s is the water concentration in air saturated at the soil surface temperature (L m^{-3}), and I_r is the infiltration rate of water through the root zone ($\text{L m}^{-2} \text{s}^{-1}$).

Many modelers calculate plant HTO concentration using a model proposed by Murphy [13], which explicitly considers contributions from both air and soil:

$$C_{pw} = [RHC_{am} + (1 - RH)C_{sw}]/\gamma, \quad (5)$$

where $\gamma = 0.909$ is the ratio of the H_2O vapor pressure to that of HTO. In deriving Eq. 5, it has been assumed that the temperature of the air, soil surface, and leaf are the same; that the mass of plant water is constant over time; and that the HTO concentration in soil water at the soil surface is equal to the concentration in root water. OBT concentrations in plants are generally calculated by assuming that the specific activity in the organic material is equal to the specific activity in the aqueous phase multiplied by an isotopic discrimination factor. More detailed OBT models would require the simulation of photosynthesis, which is too complex for inclusion in a long-term average model.

HTO concentrations in animal products (C_{aw} ; Bq L^{-1}) are usually calculated using a specific activity approach, in which the animal's total daily tritium intake from ingestion and inhalation is mixed into the total daily amount of water taken in [66, 70, 71]:

$$C_{aw} = f_{fw} C_{pw} + f_{idm} C_{OBT} + f_{dw} C_{dw} + f_{in} C_{am}. \quad (6)$$

Here C_{OBT} (Bq L^{-1}) is the OBT concentration in plant combustion water, C_{dw} (Bq L^{-1}) is the tritium concentration in drinking water and f_{fw} , f_{idm} , f_{dw} , and f_{in} are the fractions of the total water intake contributed by water in the aqueous part of food, water in the dry matter of food, drinking water, and inhalation

(including skin absorption), respectively. These fractions can be readily estimated from the diet of the animal, and so can be changed to represent different species and climates. The HTO concentration in drinking water is not easy to determine since the water could come from a surface water body of any size or from a well. Since water concentrations due to an atmospheric release are likely to be small, C_{dw} is generally set to an arbitrarily small fraction (<0.1) of C_{am} . OBT concentrations are calculated on the assumption that the specific activity in the organic material equals the specific activity in the aqueous phase, apart from isotopic effects.

Galeriu et al. [72, 73] have recently developed a metabolic approach to calculating HTO and OBT concentrations in animal products. The model explicitly takes into account transfers from HTO in the diet to HTO and OBT in the product, and from OBT in the diet to HTO and OBT in the product. The output of the model is expressed as the ratio of the concentration in the animal product to the concentration in its feed, drinking water, and inhaled air. Separate ratios are determined for HTO and OBT intakes. The model involves a number of animal and plant properties, which can be varied to reflect a given species, diet, production rate, agricultural practice, or climate.

Detailed descriptions of models that calculate long-term average tritium concentrations in the environment due to chronic atmospheric releases of HTO can be found in [62, 66, 70, 71].

Atmospheric Releases of HT

Observations As noted above, HT released to the atmosphere is not itself of radiological concern. However, HT diffuses into soil pore spaces where it is oxidized to HTO by microorganisms. Some of the HTO so formed is taken up by plants through their roots with transpiration water, and some is emitted to the atmosphere. This HTO is available for uptake by animals and humans through inhalation and ingestion. Doses from an atmospheric release of HT are determined entirely by the HTO formed in the soil.

There are very few observations of environmental HTO concentrations arising from the release of HT. Most HT sources also emit HTO directly to the atmosphere, which masks the HTO formed in the environment from the HT. The only reliable data come from an

Tritium in the Environment. Table 2 Geometric means (GM) and geometric standard deviations (GSD) for lognormal distributions for the HTO/HT ratios for air moisture, soil water, and plant water

Compartment	GM (Bq L ⁻¹ HTO)/(Bq m ⁻³ HT)	GSD
Air moisture	4	1.5
Soil water	6	1.5
Plant water	6	1.5

experiment carried out at Chalk River Laboratories (CRL) in 1994, in which an agricultural plot was continuously exposed to elevated levels of HT in air over a 12-day period [50, 74]. HTO concentrations in air, soil, and plants increased rapidly at the beginning of the release but leveled off toward the end, at which point the system was at or near steady state. The data were analyzed in terms of the steady-state ratios of HTO in air moisture to HT in air, HTO in soil water to HT in air, and HTO in plant water to HT in air [75]. The geometric means and geometric standard deviations of these ratios are shown in Table 2. The value of the air ratio in units of Bq m⁻³ HTO per Bq m⁻³ HT was 0.045.

The values of the ratios showed no dependence on downwind distance or the frequency with which the wind blew toward the receptor. The value of the plant ratio was found to apply to all plant types since the data did not show any significant difference among leafy vegetables, tubers, and fruit crops. The values of the ratios may not be directly applicable at other nuclear sites where soils are significantly different than those at CRL in terms of the properties that control HT deposition and oxidation (water content, porosity, and distribution of microorganisms).

Models The measured HTO/HT ratios explicitly account for the processes of HT deposition to soil, conversion of HT to HTO in the soil, reemission of the HTO to the atmosphere, and uptake of the HTO by plants. Thus they provide a simple way to estimate the HTO concentrations in air, soil, and plants, assuming that the HT concentration in air is known through measurement or modeling. This is the approach taken in almost all models for calculating long-term average concentrations for continuous HT sources [62, 66, 76]. The use of the ratios avoids the need to simulate the transfer processes

explicitly, which is not warranted in this type of model. Once the HTO concentrations are known, the remainder of the HT model (formation of organically bound tritium in plants, transfer to animals and humans) is the same as that for HTO. The direct transfer of HT to plants, animals, or water bodies is insignificant.

Aquatic Releases of HTO

Observations Most aquatic organisms are totally immersed in water and have high water exchange rates. Therefore, in water bodies receiving continuous input, concentrations in plant and animal tissues are in equilibrium with water concentrations. For example, in a study of a small Canadian shield lake chronically contaminated with tritium, Kim et al. [2] found that submerged plants, and the submerged parts of emergent plants, had the same concentration as the local water, whether they were rooted in the sediments or not. In the emergent parts of plants such as cattails, the HTO concentration was the average of the water and air concentrations. The HTO concentrations in fish and benthic organisms were the same as the concentration in the water and food that they accessed. Concentrations in the top few centimeters of sediment water were similar to those in the water column itself.

Because aquatic plants are immersed in an environment of uniform HTO concentration, the OBT concentration in their combustion water is the same as their HTO concentration apart from an isotopic discrimination factor, which must be determined empirically for steady-state conditions. Kirchmann et al. [77] showed that the specific activity of tritium in the organic material of aquatic plants such as algae reached 0.8 of that in the surrounding water. Lower values (0.5 for submerged plants and 0.7 for emergent plants) have been found in other studies [2, 56].

The OBT concentration in aquatic animals is also the same as the HTO concentration, reduced by a discrimination factor D_F that takes account of isotopic effects arising both in the fish and in different components of its food and water intakes. Because HTO and OBT have very different biological half-lives in aquatic animals, and because the animals can encounter different tritium concentrations as they move through their water body, it is difficult to obtain reliable estimates of D_F from animals collected in the

Tritium in the Environment. Table 3 Values of the isotopic discrimination factor for fish (D_F) obtained under equilibrium conditions

Fish type	D_F	Reference
Bullhead (catfish)	0.77	[2]
Pike	0.84	[2]
Bass	1.30 ^a	[78]
Bluegill	0.97 ^a 0.49 ^a	[79] [80]
Largemouth bass	0.41 ^a	[80]
Mosquito fish	0.34 ^a 0.67 ^a	[80] [81]
Carp	0.60 ^a	[82]
Mean	0.66	

^aThe concentrations measured in these experiments as "OBT" consisted of both exchangeable and non-exchangeable organically bound tritium. To eliminate the contribution of exchangeable OBT, the OBT/HTO ratios reported in the publications were corrected by subtracting 0.2, which represents the proportion of exchangeable hydrogen in the total hydrogen of the samples [59].

wild. More dependable values can be obtained if the animal lives its entire life in an environment where the HTO concentration does not change, or if it is exposed over a long period of time to contaminated food and water in a controlled setting. Values of D_F obtained under such conditions show a mean of 0.66 (Table 3).

Models HTO concentrations in water following chronic release to a surface water body can be calculated using standard aquatic dispersion models [83, 84]. Under the assumption of full specific activity equilibrium in all aquatic compartments, the HTO concentrations in aquatic plants and animals are both equal to the concentration in water.

Given the similarities in the processes responsible for OBT formation in terrestrial and aquatic plants, the model for OBT concentrations in aquatic plants is the same as that for terrestrial species. Accordingly, the OBT concentration in the combustion water is calculated as the product of the plant water concentration and a discrimination factor that accounts for isotopic effects in dry matter production. The plant water concentration used in these calculations should be the

concentration in the above-water part of the plant for emergent plants and in the submerged part otherwise. The model for OBT concentrations in aquatic animals takes the same form as the plant model, with a discrimination factor appropriate for animals.

Nonequilibrium Conditions

OBT/HTO ratios in terrestrial plants collected in the field near continuous tritium sources often show values greater than 1 [85–89], in apparent violation of specific activity concepts. There are a number of explanations for this, one being the significantly different formation and clearance times of HTO and OBT in plants. An HTO measurement at a single point in time reflects the instantaneous HTO concentration at that time, but a single OBT measurement reflects an integrated value over the month or two prior to sampling. If an airborne tritium plume is present when the samples are taken, the HTO concentration will be larger than the OBT concentration and the ratio will be less than 1. If the plume is not present, the HTO concentration will be less than the OBT concentration, and the ratio will be greater than 1. Since winds typically blow in any one direction at most 20% of the time at a given site, it is likely that a tritium plume will not be present at a particular sampling time, so that ratios greater than 1 will be seen more frequently than ratios less than 1.

Different translocation patterns for HTO and OBT [65], and possibly direct incorporation into plant cells of volatile tritiated organic compounds such as formaldehyde, are other mechanisms that could lead to OBT/HTO ratios greater than 1. Diurnal effects, and measurements made when the plant is under heat or water stress, may also contribute to large values of the ratio.

Large OBT/HTO ratios have also been observed in the aquatic environment. Jean-Baptiste et al. [90] found elevated tritium concentrations in fish, plants, and sediments both upstream and downstream of a nuclear generating station on the Rhone River in France. HTO concentrations in the organisms were in equilibrium with the river water, but non-exchangeable OBT levels in fish, plants, and sediments exceeded the levels in water by factors of 3, 100, and 10,000, respectively. This apparent enrichment in the organic phase was postulated to arise from the release to the river of

fine particles of tritiated compounds used in luminous paints in a clock factory further upstream. This would explain the preferential accumulation observed in sediments due to the settling of the particles in the riverbed, and to a lesser extent on plants leaves and in fish that ingest the plants. Alternatively, the large sediment concentrations may be due to the sorption of tritium from an aqueous source to organic-rich sediment grains [91].

A similar situation recently occurred in Wales, where non-exchangeable OBT concentrations in mussels and benthic fish in Cardiff Bay were found to be 1,000 times higher than the HTO concentration in the seawater. These very high concentrations were believed to arise through direct ingestion by the animals of a variety of tritiated compounds that were released to the bay in bioavailable form from a radiopharmaceutical plant [6, 92, 93]. The selective uptake of dissolved organic tritium by primitive aquatic organisms is well documented [94–96].

The above observations do not suggest that plants and animals accumulate OBT after chronic HTO exposure in either terrestrial or aquatic environments. Rather, these and similar cases represent nonequilibrium conditions for which specific activity concepts do not apply. OBT/HTO ratios less than 1 can be expected only in cases in which the OBT is formed through metabolic processes involving HTO, and in which the averaging time of the HTO and OBT concentrations is the same. However, OBT/HTO ratios measured in 22 plant samples at Thunder Bay, Ontario, which is remote from all tritium sources, varied between 3.5 and 6.2 [97]. Similarly, Baumgärtner [25] reported the accumulation of tritium in biomolecules of plants grown in tritiated water. These results are presently unexplained, but it is emphasized that OBT/HTO ratios obtained in systems where equilibrium conditions are clearly in effect are consistently less than 1 (Table 1).

Time-Dependent Concentrations Due to Acute Releases

Acute tritium releases take many forms. Accidents at nuclear power reactors or fuel reprocessing plants can result in the release of various amounts of tritium over various time intervals. Release rates generally are highly

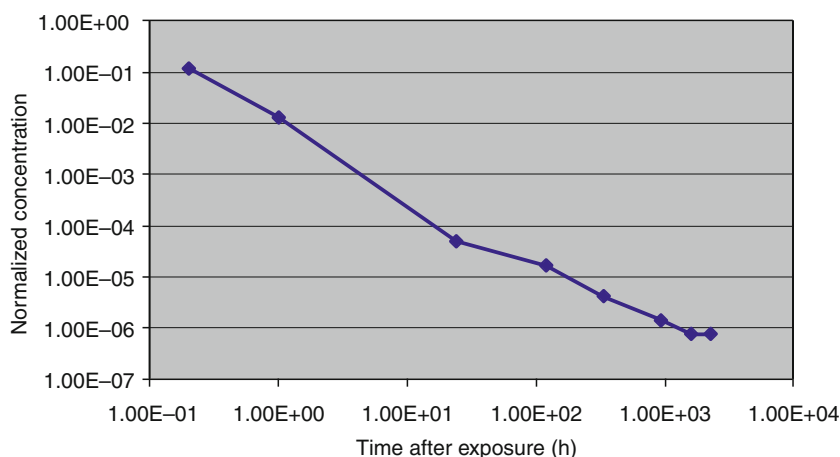
time-dependent. Livestock may undergo a step change in their dietary tritium if they are turned out on contaminated pasture after eating stored food for some time. The environmental consequences of a given acute tritium release are also scenario-specific, depending on the meteorological conditions in effect at the time of the release and shortly thereafter, and on the properties of the soils, plants, and animals that are exposed. Concentrations can be high enough in some very low probability scenarios to require protective actions to avoid human health effects.

The interest in acute releases is usually in the time-dependent environmental concentrations that result, with particular emphasis on maximum and time-integrated values. It is often difficult to obtain reliable estimates of these quantities given the unexpected nature of accidents and the typically short time available for response. As a result, almost all the reliable information on environmental tritium concentrations following a dynamic HTO release comes from controlled experiments that typically focus on a single environmental compartment (soil, plant, or animal). Since the effects of dynamic releases are difficult to generalize, they are illustrated here through a number of case studies.

Atmospheric Releases of HTO

Observations *Short-Term Exposure of Soybeans to Elevated Levels of HTO in Air:* In a series of experiments carried out by the Korean Atomic Energy Research Institute, potted soybean plants at various stages of growth were exposed for 1 h to elevated levels of tritium in a glove box [26]. The soil in the pots was covered by vinyl paper to prevent tritium from depositing to the soil, so that the air-to-leaf transfer process could be studied in isolation. Following exposure, the plants were removed from the glove box and cultivated as usual outdoors. Samples of the plant leaves, stems, shells, and seeds were taken at various times after the exposure and measured for their HTO and OBT concentrations.

Figure 1 shows the HTO concentration in the leaves and stems of a plant exposed before its pods had started to form. The observed plant concentrations have been normalized by the average HTO concentration in air moisture in the glove box during the exposure. At the



Tritium in the Environment. Figure 1

Normalized HTO concentration in soybean leaves and stems as a function of time after exposure

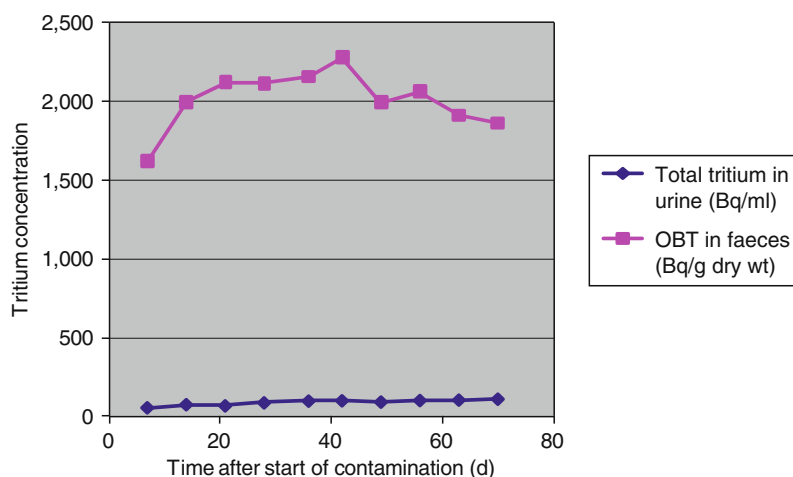
time of the first measurement, which was made 12 min after the exposure ended, the plant concentration was about 12% of the air moisture concentration. Higher values were probably achieved immediately after the exposure ended. The plant concentration dropped off by a factor of 10 in the first hour to a value just over 1% of the air moisture concentration. It continued to drop until the last two measurement points, at which time it leveled off at a value equal to the background concentration at the site. Other experiments involving plants at different growth stages showed the same general patterns but somewhat different dynamics, suggesting that tritium behavior in plants depends on the timing of the exposure relative to the growth stage. Differences in the results may also have been caused by differences in the growth rates of the plants and differences in the meteorological conditions they experienced after the exposure.

A number of exposure experiments similar to those described above have been carried out at night [14, 15]. The data show the same general features as the daytime experiments, with an increase in plant concentration during exposure and a decrease once the exposure ends. However, the rates of uptake and loss are lower at night than during the day, as are the maximum concentrations reached in the plant.

Exposure of a Sow to an OBT-Contaminated Diet: In an experiment carried out at SCK-CEN in Belgium, a pregnant sow of the Belgische Landras strain,

weighing about 180 kg, was given feed contaminated with OBT for 84 days before delivery [26]. The food had an average concentration of 577 Bq g⁻¹ dry weight and was composed of a mixture of milk powder, potato powder, and dried algae. As the pregnancy progressed, the amount of food given to the sow increased. Throughout the period, the animal had free access to water but intake was not monitored. Concentrations of total tritium (HTO plus OBT) in urine and OBT in fecal dry matter were measured every week for 10 weeks.

The results of the experiment are shown in Fig. 2. The OBT concentration in feces was much greater than the total tritium concentration in urine over the entire study period. This was expected since the tritium was administered as OBT. The fecal concentration increased rapidly in the first week of the study as the dietary OBT entered the sow's gastrointestinal tract and began to be excreted. The concentration continued to rise until the sixth week of the study and then dropped off slightly. The fecal concentration actually exceeded the dietary concentration. In the labeling process, the OBT became associated with the indigestible fraction of the feed at a higher concentration than the average for the feed as a whole, and it was the indigestible fraction that appeared as feces. The total tritium concentration in urine increased quickly at the beginning of the study, reaching a value of 57 Bq ml⁻¹ at the end of the first week. This rose to about 100 Bq ml⁻¹ by the end



Tritium in the Environment. Figure 2

Tritium concentrations in urine and feces as a function of time after the start of contamination

of the fifth week, and stayed fairly steady thereafter. HTO made up the majority of the total tritium, and arose from the breakdown of the ingested OBT during digestion. The initial rapid rise in concentration followed by a more gradual approach to steady state is typical of situations in which an environmental compartment or organism undergoes a step increase in its tritium exposure.

Models *Dynamic Compartment Models:* Models for treating the dynamic release of tritium to the environment are commonly of two types: compartment models [98–100] and process-oriented models [101–103]. The compartment models extend specific activity concepts from steady-state to dynamic conditions by considering the ratios of active-to-stable fluxes between compartments rather than the ratios of active-to-stable concentrations in the compartments. These models are formulated as a system of linear, first-order differential equations, each of which describes the change in concentration in a given compartment as the difference between inputs and losses. The simplest models have only two compartments (one organic for the whole body and one aqueous for body water) whereas more complex models can have many more. Compartment models do not simulate processes explicitly but describe them using bulk transfer coefficients, or rate constants.

Thus, the key factor in the reliability of these models is the availability of good data on which to base the rate constants. The required values are derived from data on the hydrological cycle or experimental data on tritium transfer in the environment. The latter might include tritium inventories and fluxes in a given compartment or tritium turnover times in the compartment. Biological halftimes are imposed from a simple mass balance of body water (for HTO) and carbon (for OBT).

Process-Oriented Models: Process-oriented models explicitly simulate the processes responsible for tritium transfer within and between compartments. They generally determine air concentrations using a two-step approach. In the first step, the concentration in the primary plume arising from direct release to the atmosphere is calculated using an atmospheric dispersion model. In the second step, the contribution due to reemission is taken into account using an area source approach in which the contaminated surface is divided into a number of cells. The total amount of HTO released from each cell is assumed to come from a point source in the center of the cell. Downwind concentrations due to a given cell are calculated using a dispersion model in which the plume is subject to initial widening to reflect the crosswind dimension of the cell. The total air concentration due to reemission is found by summing the contributions from each cell.

In most process-oriented models, the transfer of HTO between air and soil is modeled using an exchange velocity, v_{ex}^s (m s^{-1}), which is based on a resistance approach:

$$v_{\text{ex}}^s = (r_a + r_b + r_s)^{-1}. \quad (7)$$

Here r_a , r_b , and r_s are the HTO resistances to transfer through the air to the surroundings of the soil, through the laminar boundary layer just above the soil surface, and through the soil surface to deeper layers, respectively. All of the resistances are calculated from local data on soil properties and prevailing meteorological conditions.

Wet deposition of HTO to soil is usually calculated using a washout coefficient that depends on precipitation intensity, although more complex models are available [104]. The total tritium activity deposited by wet and dry processes on the soil is mixed into the soil water content to calculate the HTO concentration in the soil. The soil water content is determined from precipitation, evapotranspiration, and transport between layers due to matric forces. Transport is calculated using Darcy's equation [105] with appropriate values for the hydraulic conductivity and suction tension. The equations are solved numerically for each of n soil layers, where n is chosen to provide a balance between the accuracy of the model predictions and computational time.

In most process-oriented models, the transfer of HTO between air and plant is treated as a diffusion process and calculated as the product of the exchange velocity for plants, v_{ex}^p , and the HTO vapor pressure gradient between air and plant water [8, 64]. Time-dependent HTO concentrations in the plant are calculated from

$$dC_{\text{pw}}/dt = v_{\text{ex}}^p (C_a - \gamma q_s C_{\text{pw}}) / m_w, \quad (8)$$

where q_s (kg m^{-3}) is the absolute humidity of air saturated at leaf temperature (assumed equal to air temperature) and m_w is the mass of vegetation water per unit ground surface area (kg m^{-2}). The exchange velocity for plants, v_{ex}^p , is given by an equation similar to Eq. 7 but with the soil resistance replaced by a stomatal resistance, which describes transfer through the leaf epidermis into the plant. The stomatal resistance can be decreased to reflect increased uptake when the

plant leaves are wet from precipitation, or increased to model reduced uptake at night. Equation 8 can be used to estimate plant water concentrations when the HTO flux is from air to plant (when an airborne plume is present) or from plant to air (when the air is uncontaminated).

Equation 8 can be modified to account for the contribution of soil water to the plant concentration by adding a term $T C_{\text{rz}} / m_w$ to the right-hand side of the equation. Here C_{rz} (Bq L^{-1}) is the average HTO concentration in root zone soil water and T is the flux of water from the surface due to transpiration, which is given by

$$T = v_{\text{ex}}^p (q_s - q) \quad (9)$$

The reemission of HTO from soil and plants to the atmosphere can be coupled to the evapotranspiration process [106]. The flux of reemitted HTO from soil to air ($\text{Bq m}^{-2} \text{s}^{-1}$) is given by

$$F_s = -v_{\text{ex}}^s (C_a - q C_r), \quad (10)$$

where C_r (Bq L^{-1}) is the HTO concentration in soil water near the soil surface. Similarly, the flux of reemitted HTO from plant to air is given by

$$F_p = -v_{\text{ex}}^p (C_a - q C_{\text{pw}}), \quad (11)$$

F_s and F_p provide the source terms for calculating air concentrations due to reemission from soil and plants, respectively.

Plant growth and OBT formation in process-oriented models are treated using a photosynthesis model. The net photosynthesis rate is estimated from incoming solar radiation, atmospheric humidity, air temperature, soil water content, and properties of the plant canopy. The buildup of organic material is based on the amount of CO_2 assimilation [107, 108]. OBT formation is assumed to depend on the rate of organic matter production and the HTO concentration in the plant water. Some models assume that OBT can form at night. This process is generally modeled very simply by setting the nighttime formation rate to a fraction of the daytime rate, since the processes responsible for OBT formation in the dark are not well known. OBT in the green parts of plants is allowed to slowly decompose to HTO; the loss rate from grain and tubers is slower still. The plant OBT concentration is also diluted over time

by new, relatively uncontaminated dry matter that forms after the airborne plume has passed.

Galeriu et al. [109] recently proposed a process-oriented, dynamic animal model that can be applied to all mammals, both wild and domesticated. The model is based on energy need and directly relates the organic matter turnover rate to the net metabolic energy turnover rate. The relationship between these two quantities is applied both at the whole body level and at organ levels. The model requires the net energy from food as input rather than the gross energy, which is used in digestion as well as metabolism. The link between gross energy and metabolized net energy is dependent on the type of animal, its source of energy, and how it has adapted to that source.

The model includes a body water compartment to account for HTO and exchangeable OBT. The principal organic compartments were selected after considering their relevance to the food chain, radiation protection, and biological importance. Organs with high metabolic rates, including the brain and the walls of gastrointestinal organs, are combined in a viscera compartment. Blood is separated into red blood cells and blood plasma, which is the vector for the transfer of metabolites. Simplified intake compartments are included for the dynamics of their contents. Transfers from the organic compartments to blood plasma are given by the relative metabolic rate, whereas the reverse transfers are assessed using stable element mass balance.

Tritium is assumed to enter the body through the diet and is mostly absorbed from the small intestine. The dry matter intake is partitioned into a metabolizable fraction, which is associated primarily with non-exchangeable OBT and transferred to blood plasma, and an excreted fraction, which is associated with exchangeable OBT. Of the metabolizable fraction, only a part is transformed into the net input of energy and matter, the rest being lost through respiration or passed to body water.

The exchangeable fraction in the metabolizable diet is directly transferred to the body water compartment. The amount of HTO converted to OBT is calculated using a transfer rate from body water to blood plasma that is based on the fraction of organically bound hydrogen in the body derived from exchangeable hydrogen (0.25–0.35) and the assumption of mass balance in an equilibrium situation. Other losses from the

water compartment are assessed using a transfer rate based on the intake of hydrogen in drinking water, food, and respiration water.

For laboratory and some farm animals, all inputs for this model are available from the literature on animal metabolism, nutrition, and body composition. The model has recently been expanded to predict tritium dynamics in chicken broilers and eggs [110].

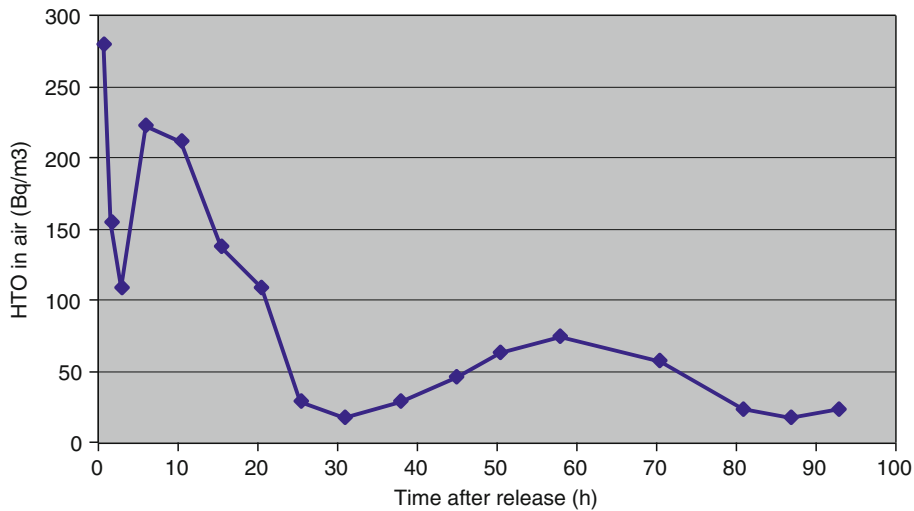
Time-dependent tritium concentrations in process-oriented models are calculated on relatively short time steps that vary in length from a few seconds to 1 h. Modeling of tritium transport on this scale requires large amounts of computer time and is not practical for calculating doses weeks or months after a short-term release. Moreover, detailed modeling beyond a few days is not necessary since adequate predictions can be obtained using simpler models and longer time steps. Accordingly, process-oriented models are used only for the first 100 h or so following a release. At longer times, dynamic compartment models running on daily time steps are used.

Atmospheric Releases of HT

Observations *Short-Term HT Release Experiments:*

The time-dependent behavior of tritium in the environment following a short-term HT release to the atmosphere has been observed in two acute release experiments, one carried out in France in 1986 [111–113] and one in Canada in 1987 [114, 115]. In the French experiment, a total of 256 TBq of HT was released over a 2-min period from a 40-m high stack beginning at 13:12 in the afternoon. The wind speed was moderate (3.5 m s^{-1}) during the release and the sky was clear. The surrounding environment consisted of cultivated fields, meadows, and small, forested areas. The soil was characterized as a silty-clay loam of fine texture. A variety of tritium measurements were carried out at a point 800 m from the stack and on two crosswind sampling lines 1,000 and 2,500 m downwind.

In the Canadian experiment, 3.54 TBq of HT were released over a 30-min period from a height of 1 m beginning at 15:20 in the afternoon. The wind was light (2.25 m s^{-1}) during the release and the weather was fair. The soil was a very fine sandy to silty loam with 2–3% organic material in the top 22 cm. The study area was vegetated with common grasses, herbs, mosses,



Tritium in the Environment. Figure 3

HTO concentrations in air for the 4-day period following the HT release in France

and large-toothed aspen saplings. Tritium concentrations were measured at various distances between 5 and 400 m from the source.

Figure 3 shows the HTO concentrations in air for the French experiment. The concentrations were measured 5 cm above ground level at a downwind distance of 800 m over the 4 days following the release, at intervals ranging from 30 min to 15 h. High HTO concentrations were observed immediately after the release as the HT was converted to HTO by microorganisms in the soil and reemitted to the atmosphere. The HTO concentrations dropped off quickly over the first 3 h as the wind carried the resuspended plume away from the sampling site, but recovered at hour 6 as the wind returned to the site. Concentrations subsequently showed an overall decrease, with occasional fluctuations depending on meteorological conditions. The Canadian results showed the same general tendencies, with a rapid buildup of HTO followed by fluctuating concentrations superimposed on a gradual downward trend.

Models The few models capable of simulating the buildup of HTO in the environment following an HT release to air are all process-oriented models. For example, in the ETMOD code [103], diffusion of HT into the soil and its subsequent oxidation is described by a first-order reaction and diffusion model based on Fick's law [105]. The reaction coefficient for oxidation is

expressed in terms of a deposition velocity, which is given by the resistance approach (Eq. 7). All deposited HT is assumed transformed into HTO. Once the soil HTO concentration is known, the rate of HTO reemission to the atmosphere is calculated using Eq. 10. HTO air concentrations due to reemission are then determined by breaking the contaminated surface into a number of cells, using a dispersion model to calculate the air concentration at the location of interest due to each cell, and summing over all cells. The ability to model reemission is crucial for an HT release, where there is no direct airborne HTO plume. Once the air concentration is known, concentrations in plants can be determined using Eq. 8.

Aquatic Releases of HTO

Observations *Concentrations in Mussels Following an Abrupt Increase in Ambient Tritium Levels:* In an experiment carried out at Chalk River Laboratories (CRL) of AECL [57], freshwater Barnes mussels (*Elliptio complanata*) were transplanted from an area in the Ottawa River with background tritium concentrations ($<10 \text{ Bq L}^{-1}$) to Perch Lake, a small lake on CRL property that has elevated tritium concentrations ($\sim 4,500 \text{ Bq L}^{-1}$). The mussels, which were all about 14–15 years old, were placed in cages at the sediment-to-water interface at a depth of approximately 0.4 m in

an area of the lake where many mussels are found naturally. Each cage was filled with 5–10 cm of sandy surface sediments originating from the area surrounding the cages, a depth that enabled mussels to position themselves in an upright position with their siphons pointed upward. Following transplantation, the mussels were sampled on an expanding time step over the course of an 86-day period. The HTO and OBT concentrations were measured in the soft tissue of each mussel sample to follow the buildup of tritium in the animals over time.

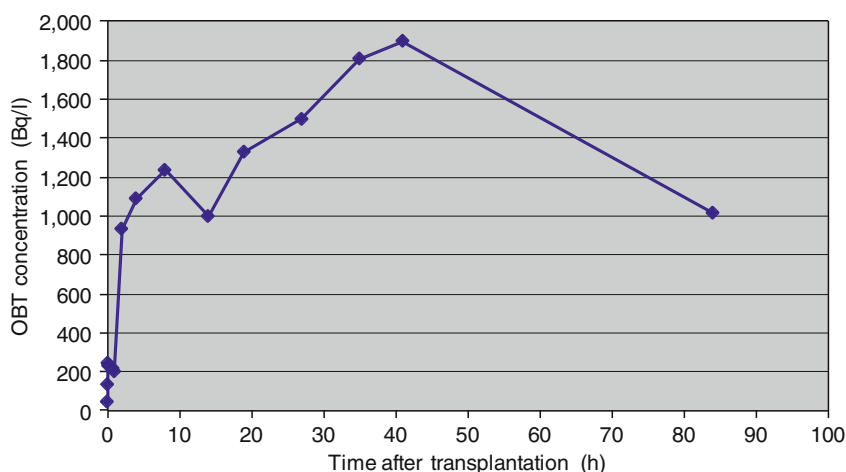
The HTO concentration in the mussels increased very quickly over time, becoming essentially identical to the water concentration within 2 h of transplantation. The OBT concentration also increased rapidly in this period, but more slowly than the HTO concentration, reaching about 25% of the water concentration after 2 h (Fig. 4). The OBT concentration continued to increase thereafter, but at a slower rate, attaining about 50% of the water concentration 40 days after transplantation. Unexpectedly, the concentration dropped by a factor of 2 between this time and the last sampling point 84 days after transplantation. This may be due to the reproductive behavior of the animals. Mussel gonads and embryos grow rapidly in midsummer and may have incorporated tritium at a faster rate than other tissues. Release of the larvae and sperm into the water column between mid-August (when the second

last mussel sample was taken) and early October (when the last sample was taken) may have caused the decline in the OBT concentration in the mussels.

Models Dynamic compartment models are usually used to simulate tritium transfer in aquatic systems subject to short-term releases. Transfers between compartments are based on published rates of tritium uptake and loss by aquatic plants and animals. Plants assimilate HTO from the water column and, for rooted plants, from sediments; animals also take up HTO from these sources and additionally from the food chain. However, the water column is by far the dominant source and HTO concentrations in plants and animals can usually be expressed in terms of the time-dependent water concentration. A single compartment representing the aqueous phase in the plant or animal is usually sufficient for this purpose. For aquatic systems subject to an increase in water concentration from an initial value of C_w^i to a final value of C_w^f , the HTO concentration (C_{HTO}) in plants and animals is given by

$$C_{\text{HTO}} = C_w^i + C_w^f [1 - \exp(-k_u t)] \quad (12)$$

Here k_u is the rate at which tritium is taken up by the plant or animal from the water column. If the water concentration decreases to zero from a value of C_w^f , the time-dependent HTO concentration varies as



Tritium in the Environment. Figure 4

OBT concentration in mussels as a function of time after transplantation

$$C_{\text{HTO}} = C_w^f \exp(-k_1 t), \quad (13)$$

where k_1 is the rate at which tritium is lost from the organism. The rate constants k_u and k_1 may be different for uptake and loss and for plants and animals, but they typically have values greater than 2 or 3 h⁻¹, implying that the HTO concentration in plants and animals comes into equilibrium with the new water concentrations in minutes or hours.

The basic mechanism of OBT formation in aquatic plants, as in terrestrial plants, is photosynthesis. Heling and Galeriu [116] proposed the following equation for the OBT concentration in phytoplankton (C_{OBT}^p):

$$\frac{dC_{\text{OBT}}^p}{dt} = 0.4 \cdot \mu \cdot f_D \cdot C_w - \mu \cdot C_{\text{OBT}}^p \quad (14)$$

Here f_D is the dry mass fraction of the phytoplankton and μ is their growth rate, which depends on light levels, water temperature, and the amount of waterborne nutrients.

For aquatic animals, OBT is formed by the metabolism of HTO in the body and by the direct incorporation of OBT taken in with food. The OBT concentration in the animal (C_{OBT}^a) is given by [117]:

$$\frac{dC_{\text{OBT}}^a}{dt} = \alpha C_f(t) + \beta C_w(t) - \kappa C_{\text{OBT}}^a \quad (15)$$

Here C_f is the OBT concentration in the animal's food, α is the transfer coefficient from HTO in water to OBT in the animal, β is the transfer coefficient from OBT in food to OBT in the animal, and κ is the loss rate of OBT from the animal.

The values of α , β , κ , and C_f all depend on the type of aquatic animal in question. The key parameter is κ , the OBT biological loss rate, which reflects both the growth rate of the animal and metabolic losses by respiration, and depends on animal mass, daily consumption rate, water temperature, and net assimilation efficiency. For zooplankton, values of κ range from 0.19 to 0.7 d⁻¹, with an average of 0.3 d⁻¹. Mollusks and crustaceans show large variability and must be considered case by case [118]. For mollusks with a weight near 1 g, $\kappa \sim 0.02$ d⁻¹; this decreases to 0.005 d⁻¹ at about 30 g. An average value for crustaceans is 0.007 d⁻¹. A special example is the mussel *Mytilus edulis*, which is taken as a reference animal for radioprotection purposes. Based on experimental data on respiration, the OBT loss rate (at 10°C) decreases

from 0.012 to 0.006 d⁻¹ as the fresh weight increases from 5 to 80 g. Similarly, κ differs by a factor of 3 for different mussel species of similar mass. For small fish, κ is in the range 0.03–0.14 d⁻¹. Data for larger fish are missing but smaller values are expected. The large range for κ demonstrates the need to choose a value specific to the animal in question in each modeling application.

For some aquatic animals such as mussels, the predictions of models with a single OBT compartment are often not in satisfactory agreement with observations. For these animals, the experimental data can be better reproduced by multi-compartment models. For example, the single OBT compartment can be broken down into two, one with a fast OBT loss rate and one with a slow rate. Alternatively, a separate compartment representing stomach contents can be added to the model to account for OBT associated with food that had not yet been digested.

Model Validation and Uncertainty

Environmental transport models are only representations of reality, and their predictions must be validated before they can be used with confidence. Most testing of environmental tritium models has taken place in international model validation programs such as BIOMOVs II (Biosphere Model Validation Study Phase II [119]), BIOMASS (Biosphere Model Assessment [61]), EMRAS (Environmental Modeling for Radiation Safety [26]), and EMRAS II (Environmental Modeling for Radiation Safety Phase II [120]), each of which included a working group on tritium. In each program, the validity of the participating models was determined by comparing their predictions with data obtained in laboratory or field studies, or with the predictions of other models. The test scenarios covered terrestrial and aquatic ecosystems and steady-state and dynamic conditions.

For most scenarios, the predictions of the various models tended to bracket the observations, suggesting that, in an average sense, the models reflect a good conceptual understanding of the environmental transport of tritium. However, the predictions of any given model differed from the observations by a factor of 2 for scenarios involving continuous releases, and by a factor of 10 or more for short-term releases.

In particular, the dynamics of OBT concentrations in plants and animals were poorly reproduced in scenarios involving short-term releases.

Galeriu et al. [109] tested the dynamic animal model described above with experimental data. In most cases, the model gave predictions that lay within a factor of 2 of the observations. An exception occurred for sheep fed a single intake of tritium-labeled glucose and acetate. In this case, the model predictions were a factor of 3–4 lower than the observations. This may have been due to the specific labeled compounds used in the study, which may not have mimicked items in the real diet. Alternatively, it may have been caused by the inability of the model to simulate the specifics of ruminant digestion of compounds in the diet. The predictions of tritium concentrations in bioassay media (urine and blood) after various intakes were of special interest, because these can be used in dosimetry. The model performed well in predicting OBT concentrations in blood after an HTO intake, but for a single OBT intake the model failed to predict OBT concentrations during the first few days. This reflects the simplified model assumption of a single metabolic pool (and a single transfer rate) for organs.

Participants in the various model validation programs also estimated the uncertainties in their predictions. For HTO endpoints in steady-state scenarios, these were roughly consistent with a 95% confidence interval (97.5th percentile divided by the 2.5th percentile) of a factor 3–4. Uncertainties were slightly higher for OBT concentrations, since the uncertainties in OBT include those for HTO plus additional ones specific to OBT itself. For scenarios involving short-term releases, the 95% confidence intervals on HTO concentrations were about a factor of 10 shortly after exposure, and larger at later times. The confidence intervals were generally smaller for OBT than for HTO, reflecting the fact that, for dynamic scenarios, HTO varies rapidly over time whereas OBT integrates.

These results are consistent with a formal uncertainty analysis of the UFOTRI code, a dynamic, process-oriented model for simulating the environmental transfer of tritium following an accidental release [121]. The calculations were carried out for a scenario involving a short-term release of HTO to the atmosphere over an agricultural area. The 95% confidence interval was estimated to cover a factor of 10 for HTO

concentrations in plant leaves and a factor of 4 for OBT concentrations in edible plant parts at harvest. These estimates reflect the uncertainty due to parameter values only and do not include uncertainties due to model structure. The uncertainties were largest for concentrations predicted immediately after exposure and decreased slightly thereafter.

The uncertainties in the models can be reduced by

- Ensuring that the air concentrations used to drive the models are of high quality and match the resolution and averaging requirements of the scenario
- Incorporating as much site-specific information as possible on land use, local soil properties, and predominant crop cultivars and animal breeds
- Basing all sub-models on the physical approaches available for the disciplines in question. For example, knowledge from the agricultural sciences should be used to improve models for crop growth, photosynthesis, translocation, and so on. Recent progress in understanding environmental hydrogen cycling should also be considered
- Recognizing and accounting for any unusual experimental conditions (water stress, an uncommon cultivar or breed) in the model application
- Using suitable spatial and temporal averaging in the models. Predicted OBT concentrations in plants and animals should be based on air concentrations averaged over the month or two prior to sampling
- Using realistic plant and animal growth models in the calculations
- Incorporating changing meteorological and environmental conditions in the models in a dynamic way
- Developing a standard conceptual model for accidental releases (an ongoing project in the EMRAS II program)

Future Directions

Future experimental work on tritium transfer in the environment should focus on the processes that are most poorly understood, namely:

- Plant uptake of HTO at night and when it is raining
- OBT formation in plants at night
- Translocation of OBT to fruit and roots
- Isotopic discrimination

- Tritium behavior in soils following deposition, including deposition of HT and conversion to HTO
- Tritium behavior in winter, including washout by snow, dry deposition to snow, and the fate of tritium in the snow pack
- The uptake and metabolism of tritium by animals

In addition, further experimental work is needed to confirm and explain previous observations in the following two areas:

- The large OBT/HTO ratios that have been observed in soils, plants and fish under conditions that are ostensibly at equilibrium
- The existence and magnitude of buried tritium

Environmental tritium models should be modified to reflect the knowledge gained through new experimental work in the areas discussed above. In addition, steady-state models for chronic releases should be revised to account for the fact that fluctuations in release rates and meteorological conditions result in a state of quasi-equilibrium in the environment, rather than the complete equilibrium assumed by the models. A rigorous sensitivity analysis of dynamic models for a variety of scenarios would be useful in establishing priorities for future work.

Finally, the uncertainty in the measurement of OBT concentrations in environmental samples, especially at low levels, is large. Certified reference materials and standard analytical procedures are needed to reduce this uncertainty and provide more confidence in the measurements.

Further work along the lines suggested above will result in better understanding of tritium behavior and concentrations in the environment, and an improved ability to estimate doses due to exposure to tritium released to air or water.

Appendix A

Definition of Organically Bound Tritium (OBT) Proposed by the EMRAS Tritium and C-14 Working Group

Definition OBT is carbon-bound and buried tritium formed in living systems through natural environmental or biological processes from HTO (or HT via HTO). Other types of organic tritium (e.g., tritiated methane,

tritiated pump oil, and radiochemicals) should be called tritiated organics, which can exist in any chemical or physical form.

Notes

1. Buried tritium is tritium that occupies exchangeable positions in large biomolecules in dry matter but that is not removed by rinsing with tritium-free water. Buried tritium therefore contributes to the OBT concentration in the traditional experimental determination of OBT. It is analogous to buried hydrogen in biochemistry.
2. OBT should not include tritium bound to sulfur, nitrogen, or oxygen (exchangeable OBT) that can be removed by washing with tritium-free water. This fraction depends strongly on the HTO concentration in effect at the time of sampling and can exchange with water vapor during analysis. Inclusion of the exchangeable fraction would lead to measurements that are highly variable and difficult to interpret.
3. From an analytical perspective, OBT is the activity in dry biomatter that is not exchangeable with water. In measuring OBT concentrations, exchangeable OBT should first be removed by moderately drying the sample without decomposing the organic molecules, washing the residue repeatedly with tritium-free water, and then drying the material again. The OBT concentration can then be determined as the tritium activity in the dry sample. This is generally done by combusting the sample and determining the activity in the combustion water by liquid scintillation counting, or by analyzing the sample by He-3 mass spectrometry. There are no generally accepted standard techniques for measuring OBT and the methods used should be documented when reporting results.
4. In the washing process, exchangeable tritium nuclei are removed and replaced by hydrogen nuclei, but exchangeable hydrogen nuclei are simply replaced by other hydrogen nuclei. Thus, measurements of OBT do not reflect the specific activity of the non-exchangeable hydrogen. This specific activity can be estimated by dividing the measured concentration by the fraction of hydrogen nuclei in the dry sample that are non-exchangeable. For example, this

fraction has been empirically determined to be 0.78 for leaf tissues, but different values may apply for other plant or animal materials. Care must be taken in comparing model predictions and experimental data that the same quantity (OBT concentration or specific activity of non-exchangeable hydrogen nuclei) is being considered.

5. OBT concentrations should be reported in units of Bq L^{-1} of combustion water. This is the fundamental unit that can be converted, if necessary, to the specific activity of the non-exchangeable hydrogen nuclei. Use of Bq L^{-1} makes it easy to compare concentrations in different media and to determine whether specific activity is depleted, preserved, or enriched when tritium is transferred from one compartment to another.
6. OBT refers to organic tritium formed from HTO by natural processes in living organisms, or in materials such as soils or lake sediments that are derived from living material. Put another way, OBT is that organic tritium found in a normal diet that imparts a dose consistent with the ICRP ingestion dose coefficient for OBT. All other types of organic tritium, no matter how they form or how they appear in the environment, should be called tritiated organics and assigned their own dose coefficient for purposes of dose calculation.

Bibliography

Primary Literature

1. Okada S, Momoshima N (1993) Overview of tritium: characteristics, sources, and problems. *Health Phys* 65:595–609
2. Kim SB, Workman WJG, Davis PA, Yankovich T (2004) Tritium concentrations in the Perch Lake aquatic ecosystem. AECL Report CTD-03700-ENA-003. Chalk River Laboratories, Chalk River
3. Diabate S, Strack S (1993) Organically bound tritium. *Health Phys* 65:698–712
4. Belot Y (1986) Tritium in plants: a review. *Radiat Prot Dosim* 16:101–105
5. Dunstall TG, Ogram GL, Spencer FS (1985) Elemental tritium deposition and conversion in the terrestrial environment. *Fusion Technol* 8:2551–2556
6. Williams JL, Russ RM, McCubbin D, Knowles JF (2001) An overview of tritium behaviour in the Severn Estuary (UK). *J Radiol Prot* 21:337–344
7. Fairlie I (1992) Tritium: the overlooked nuclear hazard. *Ecologist* 22:228–232
8. Belot Y, Ganthier K, Camus H, Caput C (1979) Prediction of the flux of tritiated water from the air to plant leaves. *Health Phys* 37:575
9. Garland JA (1979) Transfer of tritiated water vapour to and from land surfaces. In: *Behaviour of tritium in the environment*, IAEA-SM-232/74, Vienna, pp 349–358
10. Täschner M, Bunnenberg C, Camus H, Belot Y (1995) Investigations and modeling of tritium re-emission from soil. *Fusion Technol* 28:976–981
11. Hicks BB, Baldocchi DD, Meyers TP, Hosker JR, Matt DR (1987) A preliminary multiple resistance routine for deriving dry deposition velocities from measured quantities. *Water Air Soil Pollut* 36:311–330
12. Feinhals J, Bunnenberg C (1988) Laboratory investigations of HTO deposition to soils. *Fusion Technol* 14:1253
13. Murphy CE Jr (1984) The relationship between tritiated water activities in air, vegetation and soil under steady-state conditions. *Health Phys* 47:635–639
14. Atarashi M, Amano H, Ichimasa M, Ichimasa Y (1998) Deposition of D_2O from air to plant and soil during an experiment of D_2O vapor release into a vinyl house. *Fusion Eng Des* 42:133–140
15. Strack S, Diabate S, Raskob W (2005) Organically bound tritium in plants: insights gained by long-term experience in experimental and modeling research. *Fusion Sci Technol* 48:767–770
16. Murphy CE Jr (1993) Tritium transport and cycling in the environment. *Health Phys* 65:683–697
17. NCRP (1979) Tritium in the environment. National Commission on Radiation Protection report no. 62, Bethesda
18. Moses V, Calvin M (1959) Photosynthesis studies with tritiated water. *Biochim Biophys Acta* 33:297–312
19. Thompson RG, Nelson CD (1971) Photosynthetic assimilation and translocation of ^3H - and ^{14}C -organic compounds after ^3HHO and $^{14}\text{CO}_2$ were simultaneously offered to a primary leaf of soybean. *Can J Bot* 49:757–766
20. Indeka L (1981) Incorporation of tritiated water from the atmosphere into aqueous and organic components of plants. In: Jaworowski Z (ed) *Biological incorporation of tritium*. Final report to the United States Environmental Protection Agency concerning research contract no. 05-536-3, CLOR-115/D, pp 21–27
21. Diabate S, Strack S (1997) Organically bound tritium in wheat after short-term exposure to atmospheric tritium under laboratory conditions. *J Environ Radioactiv* 36:157–175
22. McFarlane JC (1976) Tritium fractionation in plants. *Environ Exp Bot* 16:9–14
23. Garland JA, Ameen M (1979) Incorporation of tritium in grain plants. *Health Phys* 36:35–38
24. Baumgärtner F, Donhär W (2004) Non-exchangeable organically bound tritium (OBT): its real nature. *Anal Bioanal Chem* 379:204–209
25. Baumgärtner F (2005) Accumulative tritium transfer from water into biosystems. *Fusion Sci Technol* 48:787–790
26. IAEA (2010) Modeling the environmental transfer of tritium and carbon-14 to biota and man. Final report, Tritium and

- Carbon-14 Working Group, IAEA EMRAS Program. International Atomic Energy Agency, Vienna
27. Ware A, Allott RW (1999) Review of methods for the analysis of total tritium and organically bound tritium. National Compliance Assessment Service technical report NCAS/TR/99/003
 28. Robertson JS (1973) Tritium turnover rates in mammals. In: Moghissi AA, Carver MW (eds) Tritium. Messenger Graphics, Phoenix, pp 322–326
 29. van den Hoek J, Kirchmann R, Juan NB (1979) Transfer and incorporation of tritium in mammals. In: Behaviour of tritium in the environment. International Atomic Energy Agency, Vienna, pp 433–444, IAEA-SM-232/74
 30. van den Hoek J, ten Have MHJ, Gerber GB (1983) The metabolism of tritium and water in the lactating dairy cow. *Health Phys* 44:127–133
 31. Hodgson A, Scott JE, Fell TP, Harrison JD (2005) Dose from the consumption of Cardiff Bay flounder containing organically bound tritium. *J Radiol Prot* 25:149–159
 32. NCAS (2001) Potential for bio-accumulation of organically bound tritium in the environment: review of monitoring data. National Compliance Assessment Service technical report NCAS/TR/2000/026
 33. Kirchmann R, Charles P, van Bruwaene R, Remy J, Koch G, van den Hoek J (1977) Distribution of tritium in the different organs of calves and pigs after ingestion of various tritiated feeds. *Curr Top Radiat Res Quart* 12:291–312
 34. van den Hoek J, ten Have MHJ, Gerber GB, Kirchmann R (1985) The transfer of tritium-labeled organic material from grass into cow's milk. *Radiat Res* 103:105–113
 35. Galeriu D, Beresford NA, Takeda H, Melintescu A, Crout NMJ (2003) Towards a model for the dynamic transfer of tritium and carbon in mammals. *Radiat Prot Dosim* 105:387–390
 36. Hill RL, Johnson JR (1993) Metabolism and dosimetry of tritium. *Health Phys* 65:628–647
 37. Rudran K (1988) Significance of in vivo organic binding of tritium following intake of tritiated water. *Radiat Prot Dosim* 25:5–13
 38. CERRIE (2004) Tritium: properties, metabolism and dosimetry. Committee Examining Radiation Risks of Internal Emitters, London
 39. Friedrich B, Schwartz E (1993) Molecular biology and hydrogen utilization in aerobic chemolithotrophs. *Annu Rev Microbiol* 47:351
 40. Ichimasa Y, Ichimasa M, Jiang H, Katsuno K (1995) In vitro determination of HT oxidation activity and tritium concentration in soil and vegetation during chronic HT release experiment at Chalk River. *Fusion Technol* 28:877–882
 41. McFarlane JC, Rogers RD, Bradley DV (1978) Environmental tritium oxidation in surface soil. *Environ Sci Technol* 12:590–593
 42. Bunnenberg C, Feinhals J, Wiener B (1986) Differences in the behaviour of HTO and H₂O in soil after condensation from the atmosphere and conversion of HT to HTO and OBT in soil relative to moisture content and pore volume. *Radiat Prot Dosim* 14:83–87
 43. Spencer FS, Vereecken-Sheehan L (1994) HTO relationships between air, soil and vegetation and seasonal dependence of HTO deposition to soil. Ontario Hydro report A-NBP-94-36-K, Toronto
 44. Dunstall TG, Ogram GL (1991) Diffusion and biological oxidation as component processes regulating the deposition of tritiated hydrogen to soils. Ontario Hydro report 90-235-K, Toronto
 45. Fallon RD (1982) Influences of pH, temperature and moisture on gaseous tritium uptake in surface soils. *Appl Environ Microbiol* 44:171–178
 46. Förstel H (1986) Uptake of elementary tritium by the soil. *Radiat Prot Dosim* 16:75–81
 47. Garland JA, Cox LC (1980) The absorption of tritium gas by English soils, plants and the sea. *Water Air Soil Pollut* 14:103–114
 48. Sweet CW, Murphy CE Jr (1981) Oxidation of molecular tritium by intact soils. *Environ Sci Technol* 15:1485–1487
 49. Wiener B, Täschner M, Bunnenberg C (1988) Tritium pathways in the atmosphere-soil system after a HT release. Niedersaechsisches Institut fuer Radiooekologie (NIR), Hannover
 50. Davis PA, Workman WJG, Amiro BD, Spencer FS, Noguchi H, Amano H, Ichimasa Y, Ichimasa M (1995) Overview of the 1994 chronic HT release experiment at Chalk River. *Fusion Technol* 28:840–845
 51. Brown RM, Ogram GL, Spencer FS (1990) Oxidation and dispersion of HT in the environment: the August 1986 field experiment at Chalk River. *Health Phys* 58:171–181
 52. Conrad R (1988) Biogeochemistry and ecophysiology of atmospheric CO and H₂. *Adv Microb Ecol* 10:231–283
 53. Spencer FS, Dunstall TG (1986) Molecular tritium conversion in vegetation, litter and soil. *Radiat Prot Dosim* 16:89–93
 54. Sweet CW, Murphy CE Jr (1984) Tritium deposition in pine trees and soil from atmospheric releases of molecular tritium. *Environ Sci Technol* 18:358–361
 55. Horton JH, Corey JC, Wallace RM (1971) Tritium loss from water exposed to the atmosphere. *Environ Sci Technol* 5:338–343
 56. Blaylock BG, Hoffman FO, Frank ML (1986) Tritium in the aquatic environment. *Radiat Prot Dosim* 16:65–71
 57. Yankovich TL, Kim SB, Baumgärtner F, Galeriu D, Melintescu A, Miyamoto K, Saito M, Siclet F, Davis PA (2010) Measured and modeled tritium concentrations in freshwater Barnes mussels (*Elliptio complanata*) exposed to an abrupt increase in ambient tritium levels. *J Environ Radioact* (in press)
 58. Galeriu D, Takeda H, Melintescu A, Trivedi A (2005) Energy metabolism and human dosimetry of tritium. *Fusion Sci Technol* 48:795–798
 59. Rodgers DW (1992) Tritium dynamics in juvenile rainbow trout, *Salmo gairdneri*. *Health Phys* 63:331–337
 60. Ostlund HG, Fine RA (1979) Oceanic distribution and transport of tritium. In: Behaviour of tritium in the environment. International Atomic Energy Agency, Vienna, pp 301–311, IAEA-SM-232/74

61. IAEA (2003) Modeling the environmental transport of tritium in the vicinity of long-term atmospheric and sub-surface sources. BIOMASS-3. International Atomic Energy Agency, Vienna
62. IAEA (2010) Handbook of parameter values for the prediction of radionuclide transfer in terrestrial and freshwater environments. Technical report series no. 472. International Atomic Energy Agency, Vienna
63. Davis PA, Kotzer TG, Workman WJG (2002) Environmental tritium concentrations due to continuous atmospheric sources. *Fusion Sci Technol* 41:453–457
64. Melintescu A, Galeriu D (2005) A versatile model for tritium transfer from atmosphere to plant and soil. *Radioprotection* 40(Suppl 1):S437–S442
65. Spencer FS (1984) Tritiated water uptake kinetics in tissue-free water and organically bound fractions of tomato plants. Ontario Hydro Research Division, Toronto, Report No. 84-69-K
66. CSA (2009) Guidelines for calculating derived release limits for radioactive material in airborne and liquid effluents for normal operation of nuclear facilities. Canadian Standard Association Standard N288.1
67. Kim M-A, Baumgärtner F (1994) Equilibrium and non-equilibrium partition of tritium between organics and tissue water of different biological systems. *Appl Radiat Isot* 45:353–360
68. Guenot J, Belot Y (1984) Assimilation of ^3H in photosynthesizing leaves exposed to HTO. *Health Phys* 47:849–855
69. Evans AG (1969) New dose estimates from chronic tritium exposures. *Health Phys* 16:57–63
70. Raskob W (1994) Description of NORMTRI: a computer program for assessing the off-site consequences from air-borne releases of tritium during normal operation of nuclear facilities. Kernforschungszentrum Karlsruhe, Karlsruhe, Report KfK-5364
71. Peterson S-R, Davis PA (2002) Tritium doses from chronic atmospheric releases: a new approach proposed for regulatory compliance. *Health Phys* 82:213–225
72. Galeriu D, Crout NMJ, Melintescu A, Beresford NA, Peterson S-R, Van Hess M (2001) Metabolic derivation of tritium transfer factors in animal products. *Radiat Environ Biophys* 40:325–334
73. Galeriu D, Melintescu A, Takeda H (2007) Risk from tritium exposure. IRPA Regional Congress for Central and Eastern Europe, Brasov. www.irpa2007romania.com
74. Davis PA, Galeriu DC, Spencer FS, Amiro BD (1995) Evolution of HTO concentrations in soil, vegetation and air during an experimental chronic HT release. *Fusion Technol* 28:833–839
75. Davis PA, Bickel GA (2000) Environmental HTO/HT ratios arising from a chronic atmospheric HT release. In: Proceedings of the international workshop on the environmental behavior and biological effects of tritium, Osaka
76. Neil BCJ (1991) An environmental pathway model for chronic emissions of tritiated hydrogen gas. Ontario Hydro report HSD-ST-91-21, Toronto
77. Kirchmann R, Bonotto S, Soman SD, Krishnamoorthy TM, Iyengar TS, Moghissi AA (1979) Transfer and incorporation of tritium in aquatic organisms. In: Behaviour of tritium in the environment. International Atomic Energy Agency, Vienna, IAEA-SM-232/96
78. Eaton D, Murphy CE Jr (1992) Tritium uptake by fish in a small stream. Westinghouse Savannah River Company report WRSC-TR-92-193-Rev 1, Aiken
79. Elwood JW (1971) Tritium behaviour in fish from a chronically contaminated lake. In: Proceedings of the 3rd national symposium on radioecology, Oak Ridge, 10–12 May 1971, pp 435–439
80. Blaylock BG, Frank ML (1979) Distribution of tritium in a chronically contaminated lake. In: Behaviour of tritium in the environment. International Atomic Energy Agency, Vienna, p 711, IAEA-SM-232/74
81. Patzer RG, Moghissi AA, McNelis DN (1973) Accumulation of tritium in various species of fish reared in tritiated water. In: Symposium on environmental behaviour of radionuclides released in the nuclear industry, Aix-en-Provence. International Atomic Energy Agency, Vienna, p 403
82. Strand JA, Templeton WL, Olson PA (1976) Fixation and long-term accumulation of tritium from tritiated water in an experimental aquatic environment. In: Watson JS, Wiffen FW (eds) Radiation effects and tritium technology for fusion reactors. Gatlinburg, TN, pp III-77–95
83. NCRP (1996) Screening models for releases of radionuclides to atmosphere, surface water, and ground. Report number 123, National Council on Radiation Protection and Measurements, Bethesda
84. Bird GA, Stephenson M, Cornett RJ (1993) The surface water model for the assessment of Canada's nuclear fuel waste management concept. *Waste Manage* 13:153–170
85. Koenig LA, Fark S, Hempelman S, Langguth KG, Pagliosa G, Papadopoulos D, Strack S (1987) Untersuchungen zum transport von tritium in der umwelt. Kernforschungszentrum Karlsruhe, Karlsruhe, Report KfK 4131
86. Hisamatsu S, Takizawa Y, Itoh M, Ueno K, Katsumata T, Sakanoue M (1989) Further study on fallout ^3H ingestion in Akita, Japan. *Health Phys* 57:559–563
87. Takashima Y, Momoshima N, Kaji T (1989) High specific activity of organically bound tritium in pine needles and search for its cause in the environment. In: Tritium radiobiology and health physics. Institute of Plasma Physics, Nagoya, pp 46–58, Report No. IPPY-REV-3
88. Brown RM (1988) Assessment of the significance of organically bound tritium in environmental materials. Atomic Energy Control Board report no. INFO-0283, Ottawa
89. Pointurier F, Baglan N, Alanic G, Chiappini R (2003) Determination of organically bound tritium background level in biological samples from a wide area in the south-west of France. *J Environ Radioactiv* 68:171–189
90. Jean-Baptiste P, Baumier D, Fourre E, Dapigny A, Clavel B (2007) The distribution of tritium in the terrestrial and aquatic environments of the Creys-Malville nuclear power plant (2002–2005). *J Environ Radioactiv* 94:107–118
91. Turner A, Millward GE, Stemp M (2009) Distribution of tritium in estuarine waters: The role of organic matter. *J Environ Radioactiv* 100:890–895

92. McCubbin D, Leonard KS, Bailey TA, Williams J, Tossell P (2001) Incorporation of organic tritium (^3H) by marine organisms and sediment in the Severn estuary/Bristol channel (UK). *Mar Pollut Bull* 42:852–863
93. Environmental Agency (2001) Potential for bio-accumulation of organically bound tritium in the environment: review of monitoring data. National Compliance Assessment Service technical report, NCAS/TR/2000/026
94. Strack S, Kistner G, Emeis C (1979) Incorporation of tritium into planktonic algae in a continuous culture under dynamic equilibrium. In: Behaviour of tritium in the environment. International Atomic Energy Agency, Vienna, p 219, STI/PUB/498
95. Strack S, Bonotto S, Kirchmann R (1980) Radioactive contamination of the marine environment: uptake and distribution of H^3 in *Dunaliella bioculatas*. *Helgoländer Meeresunters* 33:153–163
96. Strack S, Kirchmann R, Luttke A, Bonotto S (1983) Selective accumulation of organically bound tritium in the marine unicellular algae *Dunaliella bioculatas* and *Acetabularia mediterranea*. *Int J Appl Radiat Isot* 34:865–869
97. Kim SB (2010) Unpublished data. AECL, Chalk River Laboratories, Chalk River
98. Smith GM, Robinson PC, Stenhouse MJ (1995) H-3 foodchain modelling following short-term release to atmosphere. Report IE3947-1, version 2.0, Intera Information Technologies, Henley-on-Thames
99. Higgins NA, Shaw PV, Haywood SM, Jones JA (1996) TRIF, a dynamic model for predicting the transfer of tritium through the terrestrial foodchain. National Radiological Protection Board report, NRPB-R278
100. Watkins BM, Robinson PC, Smith GM (1998) Update of models for assessing short-term atmospheric releases of C-14 and tritium in the light of new information and experimental data. Ministry of Agriculture Fisheries and Food, London, Report MAFF-5044-1
101. Raskob W (1993) Description of the new version 4.0 of the tritium model UFOTRI including user guide. Kernforschungszentrum Karlsruhe, Karlsruhe, Report KfK-5194
102. Russell SB, Ogram GL (1992) ETMOD: a new environmental tritium model. *Fusion Technol* 21:645–650
103. Thompson JW, Kennedy JA, Lina JM (1992) ETMOD software quality assurance documentation, vol I. CFFTP P-9205/CFFTP-03421 (577), Canadian Fusion Fuels Technology Project, Toronto
104. Atanassov D, Galeriu D (2010) Rain scavenging of tritiated water vapour: a numerical Eulerian stationary model. *J Environ Res* (submitted)
105. Kirkham D, Powers WL (1972) Advanced soil physics. Wiley-Interscience, New York
106. Brutsaert WH (1982) Evaporation into the atmosphere. D. Reidel, Dordrecht
107. Goudriaan J (1986) A simple and fast numerical method for the computation of daily totals of crop photosynthesis. *Agric For Meteorol* 38:249–254
108. Sellers PJ (1985) Canopy reflectance, photosynthesis and transpiration. *Int J Remote Sens* 6:1335–1372
109. Galeriu D, Melintescu A, Beresford NA, Takeda H, Crout NMJ (2009) The dynamic transfer of ^3H and ^{14}C in mammals – a proposed generic model. *Radiat Environ Biophys* 48:29–45
110. Melintescu A, Galeriu D (2010) Energy metabolism used as a tool to model the transfer of ^{14}C and ^3H in animals. *Radiat Environ Biophys* (submitted)
111. Paillard Ph, Calando JP, Clerc H, Gros R, Belot Y (1988) Tritium release experiment in France: Results concerning HT/HTO conversion in the air and soil. *Fusion Technol* 14:1226–1230
112. Djerassi H, Gulden W (1988) Overview of the tritium release experiment in France. *Fusion Technol* 14:1216–1221
113. Wiener B, Täschner M, Bunnenberg C (1988) HTO re-emission from soil after HT deposition and dose consequences of HT releases. *Fusion Technol* 14:1247–1252
114. Burnham CD, Brown RM, Ogram GL, Spencer FS (1988) An overview of experiments at Chalk River on HT dispersion in the environment. *Fusion Technol* 14:1159–1164
115. Russell SB, Ogram GL (1988) Modelling elemental tritium deposition, conversion and re-emission using Ontario Hydro's tritium dispersion code. *Fusion Technol* 14:1193–1198
116. Heling R, Galeriu D (2002) Modification of LAKECO-B for ^3H . NRG report P20874/02.55114/I, Arnhem
117. Galeriu D, Heling R, Melintescu A (2005) The dynamics of tritium- including OBT- in the aquatic food chain. *Fusion Sci Technol* 48:779–782
118. Melintescu A (2009) Modelling the tritium impact on the environment. Seminar held at Universidad Politecnica de Madrid, Department of Nuclear Engineering, Madrid. <http://www.din.upm.es/drupal/files/Conf-Tritium-AMelintescu.pdf>
119. Barry PJ, Watkins BM, Belot Y, Davis PA, Edlund O, Galeriu D, Raskob W, Russell S, Togawa O (1998) Intercomparison of model predictions of tritium concentrations in soil and foods following acute airborne HTO exposure. *J Environ Radioactiv* 42:191–207
120. IAEA (2010) Environmental modeling for radiation safety – phase II (EMRAS II). <http://www-ns.iaea.org/projects/emras/emras2/default.htm>
121. Galeriu DC, Davis PA, Chouhan S, Raskob W (1995) Uncertainty and sensitivity analysis for the environmental tritium code UFOTRI. *Fusion Technol* 28:853–858

Books and Reviews

- Moghissi AA, Carver MW (eds) (1973) Tritium. Messenger Graphics, Phoenix
- IAEA (1979) Behaviour of tritium in the environment. International Atomic Energy Agency, Vienna, IAEA-SM-232/74
- (1993) Health Phys 65, Special issue on the environmental behaviour and dosimetry of tritium
- (1997) J Environ Radioactiv 36, Special issue on environmental tritium

Tritium, Health Effects and Dosimetry

KEVIN BUNDY, BERTRAND THÉRIAULT, RACHEL LANE,

JULIE BURTT, PATSY THOMPSON

Directorate of Environmental and Radiation
Protection and Assessment, Canadian Nuclear Safety
Commission, Ottawa, ON, Canada

Article Outline

Glossary

Definition of the Subject

Introduction

Physical, Chemical, and Radioactive Properties of
Tritium

Adverse Health Effects Due to Tritium

Relative Biological Effectiveness

Future Directions

Abbreviations

Bibliography

Glossary

Absorbed dose The energy deposited by ionizing radiation to a suitably small volume of matter divided by the mass of that volume. Unit gray, symbol Gy.

Beta particles High-energy negatively charged electrons or positively charged positrons that are ejected by radioactive (unstable) elements as they decay. A beta particle is identical in mass and charge to an electron. Beta particles are relatively small and can be stopped, for example, by a sheet of aluminum a few millimeters thick.

Biokinetic model A mathematical description of the behavior of radionuclides in the metabolic processes of cells, tissues, organs, and organisms. It is most frequently used to describe distribution of radionuclides among tissues and excretion.

Biological half-life Of a given radionuclide, the time required for the activity to decrease, by biological clearance, to half of its original activity.

Carcinogen Any agent, such as a chemical or a form of radiation that can cause cancer.

Chromosome Structure composed of a very long DNA molecule and associated proteins that carries

part (or all) of the hereditary information of an organism.

Chromosome aberrations Damage to chromosomes resulting in dicentric chromosomes, centric chromosomes, acentric rings, or terminal deletions.

Chronic exposure Exposure persisting in time. The adjective “chronic” relates only to the duration of exposure, and does not imply anything about the magnitude of the doses involved. Normally used to refer to exposures persisting from days to many years.

Deterministic effects Changes in cells and tissues that are certain to occur after an acute dose of radiation (in excess of a threshold value of at least 1,000 mSv), below which the radiation effect is not detected. The severity of health effects – such as skin reddening, burns, and hair loss – increase with the radiation dose received.

DNA Deoxyribonucleic acid is the molecular compound in the nucleus of a cell that forms the blueprint for the structure and function of the cell. DNA has a three-dimensional structure in which two DNA chains are held together by hydrogen bonding between the bases, forming a helix.

Dose A general term for a measure of the energy deposited by radiation in a unit mass. See the more specific terms absorbed dose, equivalent dose, effective dose, and collective dose.

Dose and dose rate effectiveness factor (DDREF) The ratio between the risk per unit effective dose for high doses and dose rates and low doses and dose rates.

Dose mean lineal energies The average or mean energy that is released by the radiant particle along its path. This value is then used to calculate the absorbed dose in Gray.

Dose rate A dose delivered over any unit of time (e.g., an annual dose is technically a dose rate).

Dosimetric model (1) For intakes of radionuclides into the body, a model that estimates the dose in various organs and tissues per disintegration of a radionuclide in a specified source organ (site of deposition or transit in the body). (2) For external exposure, model that estimates the dose rate in organs and tissues per unit activity concentration of a radionuclide in an environmental medium.

Dosimetry A scientific subspecialty in radiation protection and medical physics that focuses on

calculating the internal and external doses from ionizing radiation.

Gray (Gy) Radiation damage is dependent on the absorption of radiation energy and is approximately proportional to the concentration of absorbed energy in tissue. The gray is the SI unit of absorbed radiation dose corresponding to the absorption of 1 Joule (J) of radiation energy per kilogram of material. For gamma and beta radiations, the gray is numerically equal to the sievert.

Hematopoietic Any developmental series of cells that derives from hematopoietic stem cells and results in the production of mature blood cells.

Hereditary effect A radiation induced health effect that occurs in a descendent of the exposed person.

In vitro Latin for “within the glass,” a term used to describe experiments that are performed not in a living organism but in a controlled environment, such as in a test tube or Petri dish.

In vivo Latin for “within the living,” a term used to describe experiments using a whole, living organism as opposed to a partial or dead organism, or an in vitro controlled environment.

Internal emitter A radioactive substance that has entered the body through ingestion, inhalation, or absorption.

Isotropic effect The difference in physical properties and chemical reactivity due to the difference in mass between isotopes of the same element.

Linear energy transfer A measure of energy deposited over a distance as the energy is transferred from radiation to the exposed matter. A high value of linear energy transfer indicates that energy is deposited within a small distance (i.e., Joules μm^{-1}).

Oocyte The developing egg. It is usually a large and immobile cell.

Oogenesis Formation and maturation of oocytes in the ovary.

Orthovoltage x-rays X-rays produced by x-ray tubes operating at voltages in the 200–500 kVp range (i.e., the peak voltage), and thus possessing energy up to 200–500 keV, although there is a spectrum of energies with a peak much less than the peak tube voltage.

Relative biological effectiveness (RBE) A relative measure of the effectiveness of different radiation types in inducing a specified health effect. It is expressed as the inverse ratio of the absorbed doses

of two different radiation types, the denominator being the reference radiation that would produce the same degree of a defined biological end point.

Stochastic effects A term used to group radiation induced health effects that have a statistical risk, such as cancer or inheritable diseases. For these diseases, the probability of their occurrence increases proportionally to the radiation dose received: the lower the dose, the lower the probability of occurrence. However, at no time, even for high doses, is it certain that cancer or genetic damage will result.

Teratogenic Substances or agents that can interfere with normal embryonic development.

Tritium A radioactive isotope of the element hydrogen (symbol T or ^3H). The nucleus of tritium (sometimes called a triton) contains one proton and two neutrons. Tritium atoms can replace hydrogen atoms in water molecules to form tritiated water (HTO), in organic molecules to form organically bound tritium (OBT), in air to form tritiated gas (HT). Tritium also exists in many other compounds such as tritiated hydrocarbons, tritiated particulates, tritiated thymidine, and metal tritides (tritium bearing metals).

X-ray Ionizing electromagnetic radiation emitted by an atom when it has been bombarded with electrons. X-rays differ from gamma rays in that they are emitted from the orbiting electrons, not the nucleus, and they have a much wider energy range, or spectra. A “soft X-ray” is one of low energy, generally below 100 keV.

Definition of the Subject

Tritium is a radioactive isotope of hydrogen that is produced both naturally and by man-made activities such as nuclear power plants. As a hydrogen isotope, a tritium atom behaves in the same physical and chemical manner as hydrogen atoms and so it can become part of water molecules and organic molecules. In these forms, tritium is very mobile and can enter biological systems and has the potential to damage living cells.

The biological damage caused by tritium radiation is the same as that from other radiations, but properties such as being an internal risk, being a very weak beta emitter, and being part of organic molecules

distinguish its effects from that of other radiations such as x-rays and gamma rays.

Tritium is also distinguished in that it is one of the most studied radioisotopes with respect to causing biological effects. This contribution reviews what is known of the biological effects of tritium, how its radiation compares in effectiveness with other radiations, and how the dose of tritium is assessed in humans.

This present report is an abridged version of the CNSC report “Health Effects, Dosimetry and Radiological Protection of Tritium” which was published in March 2010 and is available at <http://nuclearsafety.gc.ca>.

Introduction

Tritium (T or H-3), hydrogen’s only radioactive isotope, can exist in several forms including tritiated water (HTO), tritiated gas (HT), and organically bound tritium (OBT). The emitted beta particle from tritium has a relatively low mean energy causing it to have a very short range in dry air. It can be absorbed by sheets of plastic, glass, or metal and cannot penetrate the dead layers of skin. Adverse health effects from tritium beta radiation therefore can only occur if inhaled, ingested, or absorbed into the body through the skin in large quantities. Once taken into the body, the beta particle possesses sufficient energy to ionize atoms and molecules.

This present report is an abridged version of the Canadian Nuclear Safety Commission’s report “Health Effects, Dosimetry and Radiological Protection of Tritium” which was published in March 2010 as part of the CNSC’s Tritium Studies Project. This 3-year research and information-gathering project was initiated as a result of questions raised during a public hearing related to the licensing of a tritium facility. The goal of the research was to expand the body of knowledge on tritium and to further enhance regulatory oversight of tritium-related activities in Canada.

The objective of this report is to provide an overview of the current knowledge of what is known about the adverse health effects from tritium radiation. To understand why tritium poses a health risk, a basis of tritium’s physical, chemical, and radiological properties is required and is

provided in section “[Physical, Chemical, and Radioactive Properties of Tritium](#).” This includes an introduction to the organic forms of tritium collectively known as OBT.

The section on “[Adverse Health Effects due to Tritium](#)” discusses the adverse health effects from tritium of which there are two main types: deterministic and stochastic. The discussion is primarily devoted to laboratory studies, but a brief overview of relevant human epidemiological studies is provided as well as a review of what is known about the biological effects of OBT. In addition, section “[Non-radiological Effects of Tritium](#)” discusses the theoretical effects from tritium that are caused not by its radiation, but instead by the form it changes into after disintegration.

The effectiveness of a particular radiation producing a specific biological effect is often determined through comparison to the dose required to produce the same effect by a standard radiation such as gamma rays. This is conventionally referred to as the “relative biological effectiveness” (RBE). The section on “[Relative Biological Effectiveness](#)” gives an overview of a number of studies that have measured the RBE of tritium beta radiation and offers a possible explanation in section “[RBE Literature Reviews and Experimental Studies](#)” on why tritium has a higher RBE than other radiations with much higher energies.

Finally the section on [Dosimetry and Biokinetics](#) provides a review of the biological modeling which describes how tritium and OBT behave in the body and how the resultant dose from that distribution is calculated. This includes discussions of intakes by inhalation, ingestion, and skin absorption and dose models for pregnancy and nursing.

Physical, Chemical, and Radioactive Properties of Tritium

Hydrogen, the simplest and most common element, has three isotopes:

- Protium (H-1), which contains only a proton in its nucleus and is by far the most common hydrogen isotope
- Deuterium (H-2), which contains a proton and a neutron
- Tritium (H-3), which has a proton and two neutrons, and is hydrogen’s only radioactive isotope

Tritium is produced:

- Naturally through various interactions of cosmic rays and atoms in the atmosphere – the principal one being the neutron irradiation of nitrogen
- As a by-product of the operation of nuclear and research reactors including the capture of a fission neutron by a deuterium atom or neutron capture by Boron-10 and Lithium-6

Radioactive Properties

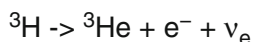
Tritium decays to a stable form of helium by emitting an electron from its nucleus (beta radiation) and an antineutrino (see Fig. 1):

It has a radioactive half-life of 12.3 years and a specific activity of 3.56×10^{14} Bq/g. The emitted beta particle has a mean energy of 5.7 keV and a maximum energy of 18.6 keV. Despite this very low energy, these particles possess sufficient energy to ionize atoms and molecules. Due to its very low energy, tritium's beta radiation has a very short range (approximately 6 mm) in dry air; is completely absorbed by sheets of plastic, glass, or metal; and cannot penetrate dead layers of skin. However, this beta radiation can pose a health risk if tritium is taken into the body. The maximum range of a tritium beta particle is approximately 6 μm in living tissue [1], compared to typical cells (or cell nuclei) that have diameters from about 7–30 μm (nuclei of about 6–15 μm) [2].

Chemical Forms and Properties

Tritium is almost chemically identical to the other hydrogen isotopes and can exist in several chemical forms including: HTO, HT, and OBT.

Tritiated Water The most common form of tritium is HTO, where a tritium atom replaces a hydrogen atom in water (H_2O) to form HTO. HTO has the same chemical properties as water and is also odorless and colorless. Water with a tritium activity of 1 Bq/L contains less than 1 tritium atom per 100,000 million,



Tritium, Health Effects and Dosimetry. Figure 1
Decay formula for tritium

million (1 in 10^{17}) molecules. The majority of tritium in the atmosphere is in the form of HTO and can be transferred to humans by inhalation, skin absorption (liquid and vapor) [3, 4], or ingestion of drinking water or food [5, 6]. HTO exposure is generally the most important consideration in assessing dose, since HTO quickly reaches equilibrium with water in the body and is distributed uniformly to all soft tissues. The International Commission on Radiological Protection (ICRP) [5] recommended that internalized HTO be assumed to be completely and instantaneously absorbed and distributed uniformly with all body water. As a result, the concentration in sweat, sputum, urine, blood, perspiration, and exhaled water vapor is taken to be the same [7]. HTO is excreted via urine, feces, sweat, and breath [8].

Tritiated Gas HT is formed when a tritium atom replaces a hydrogen atom to form a tritium–hydrogen bond. In its elemental form, HT is an invisible, odorless gas chemically identical to hydrogen gas. HT is relatively inert in biological systems and has a very low uptake into body fluids and tissues [7]. The main exposure pathways of HT include inhalation or skin contact with HT-contaminated surfaces. Releases from tritium processing facilities (such as self-luminous light manufactures, tritium recovery facilities, and nuclear fuel processing facilities) represent the primary source of exposure to HT. HT can be oxidized in the atmosphere to HTO. If deposited on the ground, it is converted to HTO through microbial agents near the soil surface [9].

Organically Bound Tritium Following an intake of tritium (typically in the form of HTO) by plants or animals, a fraction of the tritium can become incorporated into organic molecules such as carbohydrates, fats, or proteins and is termed organically bound tritium (OBT; [note that OBT is an acronym that stands for organically bound tritium and is not a chemical formula such as HT and HTO]). Within the body, OBT can become incorporated into various compounds such as amino acids, sugars, and structural materials such as collagen [10, 11]. OBT can also enter the body directly by ingestion of tritiated food, by inhalation of volatile organic vapors or aerosols, or it can be formed in vivo from tritium that is present in the general body

pools after exposure to other tritium-containing compounds [12]. If the tritium atom is attached to a carbon atom in an enzymatically catalyzed reaction, it is essentially fixed to that molecule until the molecule is metabolized. Conversely, tritium atoms bonded to oxygen, sulfur, nitrogen, or phosphorous atoms will readily exchange with hydrogen atoms in the surrounding cellular water and are therefore considered as an exchangeable tritium compound [12]. While both forms can be regarded as OBT, it is the nonexchangeable carbon-bound tritium that exhibits longer retention time in the body characteristic of organic molecules. Tritium–hydrogen exchange reactions occur in organic molecules with chemical groups such as nitrogen monohydride (NH), hydroxyl (OH), and thiol (SH). Nonexchangeable tritium is incorporated into more stable bonds with carbon atoms by enzyme catalyzed reactions [12].

Adverse Health Effects due to Tritium

Tritium beta radiation does not penetrate the dead outer layer of the skin. Therefore, tritium only poses a health risk if inhaled, ingested, or absorbed into the body through the skin in large quantities. If the chemical form is HTO, it will not collect in any specific tissue or organ but will distribute itself uniformly throughout the body. If it is part of an organic molecule, it may be incorporated into specific molecules or tissues. Generally speaking, the proximity of tritium to key molecules – principally deoxyribonucleic acid (DNA) – will determine the importance of the damage it will inflict when it undergoes nuclear transformation and emits a beta particle. Since HTO and most forms of OBT are assumed to be uniformly distributed in body tissues and cells, one would expect intakes of tritium to result in radiation effects similar to that of whole-body exposures from other types of low-linear energy transfer (LET) radiation, such as x-rays and gamma rays. Of course, the severity of the damage would vary according to differences in the RBE of the different radiations as well as to dose and dose rate effects.

An exception to these direct interactions is a phenomenon known as the “bystander effect,” where nonirradiated cells exhibit the same symptoms as nearby irradiated cells. This has been demonstrated in many studies [13], using primarily alpha particle and

photon irradiation but the effect has also been demonstrated in a study using tritium [14]. Most of the studies have been conducted in vitro, but the bystander effect has been observed in a limited number of in vivo studies and it is thought to be due to an unidentified chemical messenger. What is known is that the bystander effect occurs at the cellular level of biological damage and any damage induced by it will have already been accounted for in dose–response assessments.

As with all other forms of ionizing radiation, adverse health effects from tritium exposure can be classified as deterministic or stochastic.

- Deterministic effects:
 - Occur within relatively short time frames after exceeding a dose threshold
 - There is an increase in severity (e.g., cataracts, skin reddening) as dose increases
- Stochastic effects (cancer and hereditary effects):
 - Are assumed not to have dose thresholds
 - Have a greater probability of occurring as dose increases

Sections “[Deterministic Effects](#)” and “[Stochastic Effects](#)” review several studies looking at deterministic and stochastic effects, respectively. The review of deterministic effects includes lethality (death), teratogenicity (birth defects), and reproductive effects. The review of stochastic effects includes carcinogenicity (cancer) and hereditary effects. Section “[Human Epidemiology](#)” reviews epidemiological studies on workers and in the public that have been exposed to tritium.

Note that the conversion of an intake of tritium in becquerels to a dose is dependent on a number of factors such as the assumed turnover rate, or biological half life, of body water. This factor can differ by a factor of two or more between studies and consequently, the final calculated dose for the same tritium intake may also vary by a similar factor. Where no dose information is given, or if the cited literature was dated (e.g., 1950s), the doses were calculated for this review. The dose calculations assumed an HTO half-life in mice of 1.2 days which is consistent with that of Rodgers [15] and Richmond et al. [16].

Deterministic Effects

Lethality Brues et al. [17] demonstrated the lethality of tritium by inducing death in mice by administering

one-time tritium injections of 0.126–8.4 GBq. The researchers reported a lethal dose ($LD_{50/30}$) that would kill 50% of the population within 30 days of 37 MBq/g of body weight (BW). Their mice had an average weight of 20 g, which translates to an initial intake of about 1.11 GBq. Brues et al. [17] calculated the dose from this intake to be about 1,000 “reps” (older unit, equal to about 9 Gy). Using modern dosimetry calculations, it appears that a dose of 37 MBq/g BW gives a dose of around 6 Gy, due to difference in assumed water biological half-lives. Furchner [18] determined a $LD_{50/30}$ value of 8 Gy, by a single injection of about 24 MBq to female mice, but again, this translates to a dose of about 4 Gy with newer biological half-life. Yokoro et al. [19] (see also Yamamoto et al. [20]) found a $LD_{50/30}$ of about 8 Gy from a single injection of 0.56 GBq in one strain of mice and 0.93 GBq injection in another.

Yamamoto et al. [20] investigated chronic intakes of HTO that would be lethal in mice by continuously administered HTO. They found that the concentration of tritium reached a plateau in organs and blood after about 7 days. A typical time of 2 weeks for hematopoietic death occurred following intakes of drinking water with tritium concentrations from 148 to 592 GBq/L. The lowest total dose inducing death was 11 Gy at a drinking water concentration of 148 GBq/L. All observed deaths were due to hematopoietic failure.

Although details are vague, tritium was reported to be the cause of death of two workers in 1953 in Chelyabinsk-40 (now Ozyorsk) Russia. Doses were estimated at 10–12 Gy, indicative of massive intakes [21, 22].

In summary:

- The radiation dose from tritium necessary to cause death in mice is slightly greater than an acute external irradiation by x-rays or gamma rays, with a tritium $LD_{50/30}$ of about 6–9 Gy. This corresponds to an acute intake in the order of 1 GBq or a chronic dose of about 11 Gy (~ 0.5 GBq) when tritium is orally administered.
- An unusual finding was that hematopoietic death prevailed over the gastrointestinal syndrome following the exposure to radiation from tritium. The latter would have occurred if a similar dose had been given with gamma or x-rays [20].

Teratogenic Effects Teratogenic effects occur when an agent interferes with the development of an embryo or a fetus in utero. Teratogenic effects are not hereditary; that is, they are not passed on to future generations. In general, a tissue’s sensitivity to radiation is directly proportional to the rate of proliferation of its cells. Therefore, a fetus that is rapidly developing from an embryo is expected to be more sensitive to radiation than an infant, child, or adult. This inference is supported by the results of experiments in animal models and by experience with human populations exposed to doses of radiation above 100 mSv (for example, medical exposures, atomic bomb survivors). Radiation’s most significant harmful effects to the human fetus include mental retardation, delayed intrauterine growth, and childhood cancers (such as leukemia) [23, 24].

Straume and Carsten [25] reviewed current literature looking at tritium exposure during fetal development. They pointed out that some fetal cells rapidly divide during certain stages of development and differentiate to form organs and tissues while others such as neurons or oocytes involve very little or no further cell proliferation. Consequently, tritium that is incorporated into “long-lived” molecules (becoming OBT) could result in large integrated doses over the lifetime of the cells, since the tritium would not be diluted by further cell proliferation. The dose from tritium incorporated into molecules such as DNA and histones of rapidly dividing cells was small compared to the dose from HTO as observed by Commerford et al. [26]. Conversely, for long-lived cells, the dose from tritium incorporated into DNA and histones could exceed that from HTO. This was also discussed in some depth by the Advisory Group on Ionizing Radiation (AGIR) reporting to the UK Health Protection Agency [27] and will be presented in the following discussion on reproductive effects. The AGIR [27] concluded that tritium could be incorporated during pregnancy into the DNA of fetal oocytes and remain there until fertilization decades later; however, it also concluded that DNA is not a stable molecule and that there is a turnover of its molecules over time. The net radiological impact of this phenomenon is unclear, but would be related to dose and the amount of tritium deposited within the DNA of the long-lived cells. To be deposited within DNA, tritium would have to be bound to

a specific organic molecule (such as a nucleic acid) that could become part of a DNA molecule. All OBT compounds represent a fraction ($\sim 13\%$) [28] of the total tritium in the environment, but OBT molecules could be any one of the thousands of different types of organic molecules. It is therefore reasonable to expect that a low number of OBT compounds are used to build DNA. The section of the report on reproductive effects (see section “[Reproductive Effects](#)”) discusses this phenomenon in more detail.

Dekaban [29] first summarized radiation-induced teratogenic effects in humans when he reported developmental effects due to medical x-ray treatments in the 1920s and early 1930s, when radiation was widely used to treat many diseases. The report evaluated a group of children who were exposed in utero to radiation when their mothers had pelvic x-ray exposures. Doses of 200–300 R ($\sim 2\text{--}3$ Gy) consistently induced fetal damage. The following were the most frequent anomalies identified in children who were irradiated during intra-uterine life: (1) small size at birth and markedly stunted growth, (2) microcephaly, (3) mental retardation, (4) microphthalmus, (5) pigmentary degeneration of the retina, (6) genital and skeletal malformations, and (7) cataracts. However, Mulvihill et al. [30] later reported that these effects depend on gestational age and that other case reports of radiotherapy during pregnancy showed that anomalies did not always develop.

Zamenhof and Van Marthans [31] studied how five generations of rats were affected by pre- and postnatal exposure to HTO. Female rats were given HTO (111 MBq/L) beginning in adolescence and throughout pregnancy. This exposure to tritium did not produce any signs of radiation illness or increase in cataract formations in the mothers. The courses and outcomes of pregnancy were also normal, but 60% of the newborn rats exhibited hematomas, edemas, and subdural hemorrhages. None of these effects lasted beyond 30 days of age.

A study by Laskey et al. [32] maintained rats on activities of 0.37–370 kBq HTO/mL of body water from conception of the first generation until delivery of the second generation. The corresponding dose rates were 0.03–30 milligray per day (mGy/day). An exposure to 370 MBq/L of body water resulted in a 30% weight reduction of the testes in the first generation of adult males, but there was no impairment of growth or

reproductive ability. The following were noted in the second generation of newborns: a relative reduction of brain weights at exposures of 3.7–370 MBq/L (0.3–30 mGy/day); a decrease in BW at 37 MBq/L (3 mGy/day) and 370 MBq/L (30 mGy/day); a decrease in litter size and an increase in resorption at 370 MBq/L (30 mGy/day). In their discussion, Laskey et al. [32] noted that similar effects had not been observed in other studies with comparable doses of Cobalt-60 gamma radiation. Laskey et al. [32] did not observe litter size effects or resorption either at or below 37 MBq/L (~ 3 mGy/day), although higher dose rates showed statistically significant reductions in the weights of both male and female rats.

Bursian et al. [33] looked at effects on rats continuously exposed to HTO concentrations of 0, 37, 370, and 3,700 Bq/mL from conception to birth. The in utero exposure to doses as low as 0.66 Gy (370 MBq/L) produced measurable and persistent decreases in brain weight and increases in norepinephrine concentrations at 21 and 45 days postnatal. No differences from controls were observed in the rate of turnover or the concentrations of dopamine, acetylcholinesterase, or monoamine oxidase; although it is unclear what effects would have occurred at different concentrations.

In an experiment by Jones et al. [34], drinking water containing HTO was given to pregnant squirrel monkeys at levels ranging from 16 to 1,000 times the “permissible level for human consumption.” No effects in newborn progeny were seen in terms of BW, body dimensions, organ weights, hematological patterns, the histology of organs, or tissues, with the exception of the ovaries. Dobson et al. [35, 36] found a decrease in the number of immature oocytes at the highest doses.

In summary, tritium radiation-induced teratogenic effects have been demonstrated in laboratory studies and are consistent with similar effects from external photon radiation. Teratogenic effects in animal studies start to occur at doses of about 0.4–0.6 Gy from chronic intakes of HTO.

Reproductive Effects Reproductive organs and tissues in both males and females are among the most radiosensitive tissues, with mature gametes being the most sensitive cells. In the male, spermatogenesis is a proliferative and continuous process, where sperm

originate from spermatogonial stem cells and undergo many developmental stages, from the primordial germ cell to the mature spermatozoa. This process continues throughout adult life and due to continual replenishment, testes are among the most radiosensitive tissues.

In the female, germ cells are called oocytes. They are produced in the ovary and give rise to the ovum (egg), which can be fertilized. Oocyte formation (oogenesis) occurs in the female during fetal life, with primary oocytes produced in the ovaries of the fetus by the fifth month of pregnancy. About seven million primary oocytes are initially produced but only about 400,000 remain by puberty. The oocyte development process includes several stages of maturation, from the production of oogonia from primitive germ cells to development of primary oocytes and formation of definitive oocytes at puberty.

Effects in the Female Since tritium distributes itself throughout the body, it can be taken up by a developing oocyte and incorporated in its DNA. Tritium incorporated into oocyte DNA could theoretically irradiate the oocytes over a period of 30 or more years. Because oocytes do not divide until fertilized, there is little turnover of the DNA molecules which implies that the biological half-life of the tritium imbedded in oocyte DNA could approach the radioactive half-life of tritium, 12.3 years. The AGIR [27] points out that while a biological process gradually exchanges all the components of DNA molecules – including that of tritium-labeled DNA – it would take 50 years to replace 2–5% of a cell's genome (DNA). Therefore, most of the tritium incorporated in an oocyte will remain there for its whole life.

Straume and Carsten [25] reviewed tritium effects on oocytes, reporting that most of the information on radiosensitivity in humans came from autopsies of women who had been exposed to substantial doses of external radiation [37], and from fertility histories of women exposed to radiation from therapy or atomic bombs [38–40]. In all cases, the data were for adult women.

Available data for human females suggest that in adult women, maturing and mature oocytes are more sensitive than immature ones. This is in contrast with observations of mice, whose immature oocytes are by far the most sensitive to radiation killing – about

100 times more sensitive than matures ones [41]. In women, an exposure of 2.5–6 Gy will lead to permanent sterility (effect on resting oocytes) [42, 43]. The threshold for temporary effects on fertility appears to be at least 0.6 Gy of acute x-rays and 1.5 Gy for protracted or fractionated low-LET radiations [42].

In studies on both mice and rats, Satow [44, 45] looked at tritium effectiveness in killing oocytes. While the study primarily examined the RBE of tritium, significant oocyte reabsorption was found at HTO intakes of 0.17 MBq/g of BW, corresponding to a total dose of 39 mGy [41].

To assess the risk that tritium radiation would damage oocyte DNA, the AGIR [27] modeled the dose and hereditary risk in a critical group of pregnant women, who consumed an annual total of 24 kg of tritium-contaminated flounder from Cardiff Bay. It was assumed that the flounder had a tritium concentration of 50 kBq/kg, which implies that the critical group ingested 1,200 kBq of OBT per year. At this rate of intake, the tritium in the body would reach an equilibrium concentration of 175 kBq. The model estimated that approximately 4% of oocytes could experience a tritium disintegration within the next 30 years following intake and that the frequency of severe hereditary effects resulting from this would be expected to be around 1 in 1,000,000 (10^{-6}). This is comparable to the 3–4% rate of spontaneous birth defects that are observable and inheritable.

Overall, female reproductive systems – and principally those of mice – appear to suffer the greatest adverse health effects from tritium exposure. As Baker [41] reports, the mouse oocyte is about 100 times more radiosensitive than the human oocyte; therefore doses that affect mice oocytes would not be expected to affect human ones.

Effects in the Male Unlike oocytes, spermatogonia are continuously produced from stem cells throughout adult life. Like all tissues that are rapidly replaced, there are certain germ-cell stages that are highly sensitive to cell killing by ionizing radiation. Experiments in mice have shown that the most sensitive cells are the type A and B spermatogonia, which can be reduced by 50% by doses of only about 0.3 Gy of acute x-rays [46, 47] (as reported in Straume and Carsten [25]). The spermatid and spermatozoa stages are much less sensitive

than the spermatogonia stage [48]. Lambert [49] found a 27% reduction of mouse spermatogonia that were injected with tritiated thymidine (3HDtr) at a dose of 185 kBq/g of body mass and with HTO at a dose of 2,220 kBq/g of body mass (dose to the nucleus of 84 mGy and 49 mGy, respectively).

Although the results are from a single individual, UNSCEAR [40] reported temporary sterility after doses of only 0.15 Gy of acute x-rays and permanent sterility after acute doses of 3–5 Gy. Chronic exposure studies in dogs demonstrate that x-ray dose rates of 1–2 mGy/day do not impair sperm production [43]. If this data were transposed to humans it would imply that the threshold for reduced sperm production in the adult male may be in the order of 10 mGy/day. For continuous exposure, this would add up to 3–4 Gy/year [25]. A dose of 10 mGy/day due to tritium would require an intake of about 500 MBq/day.

Stochastic Effects

Carcinogenicity Carcinogenicity is the ability of a radiological, chemical, or biological agent to produce cancer. Ionizing radiation was one of the first such agents observed to induce cancer from its early uses in the 1900s. Many of the first researchers were unaware of the risk that radiation presented and succumbed to radiation-induced cancer. The study of populations exposed to radiation, such as the atomic bomb survivors, medical patients, and nuclear energy workers, has provided the basis of our current radiation dose limits. Most of the populations studied were those exposed to external radiation, such as x-rays, gamma rays, or neutrons. Risks to humans from exposure to internal emitters are largely limited to uranium miners exposed to radon progeny, radium dial painters, and patients injected with thorotrast [50]. While epidemiological studies are ongoing, to date no human studies have ever demonstrated tritium-induced cancer (see the section on human epidemiology).

Due to a lack of direct experimental evidence, the cancer risk posed by internal emitters to humans has largely been derived from populations that received high, external exposures such as those in the Japanese atomic bomb survivors Life Span Study [51]. Many studies have looked at the incidence of cancer among laboratory animals exposed to gamma radiation and

x-rays. While it did not involve exposures to tritium, one of the most comprehensive studies on the effects of radiation was done by Tanaka et al. [52], who looked at cause of death and neoplasia in 4,000 mice exposed to Cesium-137 gamma rays at 21, 1.1, and 0.05 mGy/day. Equal numbers of male and female mice were exposed for approximately 400 days with resultant doses of 8,000, 400, and 20 mGy. Compared to the nonirradiated controls, there was a significantly higher frequency of myeloid leukemia in males, of soft tissue neoplasms and malignant granulosa cell tumors in females, and of hemangiosarcoma in both sexes exposed to 21 mGy/day (8 Gy cumulative dose). The number of multiple primary neoplasms per mouse was also significantly increased in mice irradiated at 21 mGy/day. Both sexes exposed to 21 mGy/day had significantly shorter life spans than control groups, but only the females had significant shorter life spans when exposed to 1.1 mGy/day (0.4 Gy). Females exposed to 0.05 mGy/day (0.02 Gy cumulative dose) also had a shorter average life span by about 8 days, but this was not statistically significant.

Many laboratory studies on animals have clearly demonstrated that tritium can induce cancer, although this has not been studied as extensively as gamma and x-ray exposures.

In Yamamoto et al. [53], HTO was administered to mice in drinking water at concentrations that ranged from 9.25 GBq/L (0.240 Gy/day) to 0.37 GBq/L (0.010 Gy/day). Female mice maintained on this drinking water survived for more than 150 days, but 70–80% of them developed tumors. In the range from 1.85 to 9.25 GBq/L of HTO (0.048–0.240 Gy/day), the main cause of death was thymic lymphoma. However, at 0.925 GBq/L (0.024 Gy/day), the incidence of thymic lymphoma was lower, while tumor incidence was greater. In addition, tumor types became more diverse at lower concentrations of HTO. The latent period of tumor development was shorter and the life-shortening effect was more marked in this tritium beta irradiation study than in other studies that used x-ray or gamma irradiation. However, Yamamoto et al. [53] did not attempt to estimate the RBE of the tritium exposure.

Yamamoto et al. [54] administered tritiated drinking water to mice throughout their lives, resulting in dose rates of 0.2, 0.9, and 3.6 mGy/day. The thymic

lymphoma dose–response curve started to increase exponentially at about 0.9 mGy/day, giving a threshold at about 12 mGy/day (although there was no significant difference from controls at 0.9 mGy/day). The group receiving 3.6 mGy/day had significantly shortened life spans compared to controls, while the other groups did not. Increases in BW were noted for the animals that received lower dose rates (0.9 mGy/day). Yamamoto et al. [54] concluded that the threshold dose rate for tumor induction by HTO is around 12 mGy/day, and that the threshold dose rate for Cobalt-60 gamma irradiation is higher than that for tritium beta irradiation.

Gragtmans et al. [55] estimated the RBE of tritium beta radiation to induce tumors in the breast tissue of rats. Incidence of mammary tumors over time was compared to that of controls in a rat strain where these tumors occur naturally. Tritium was injected intraperitoneally at concentrations of 45–370 MBq/100 g BW followed by four additional injections at 2-day intervals, producing doses of 0.46, 0.92, 1.63, and 3.85 Gy. The reference radiation was 200 kVp x-rays delivered over 10 days in doses of 0.29, 0.57, 1.1, and 2.0 Gy. There was early onset of the tumors at 150 days posttreatment for all irradiations when compared to controls. Since these animals would naturally develop mammary tumors, the irradiation likely decreased the onset time of the tumor but did not necessarily increase the overall incidence. Therefore, some care must be taken when interpreting the results.

Johnson et al. [56] estimated lifetime incidence of myeloid leukemia in seven groups of ~750 mice given single injections of HTO of 90, 180, or 270 MBq each, corresponding to doses of 0.85, 1.86, and 3.04 Gy. The incidence of myeloid leukemia increased from 0.13% (control) to 6–8% (treated) with apparent excesses at the lowest doses. The calculated RBE for tritium beta rays was 1.0 ± 0.5 to 1.3 ± 0.3 (best estimate 1.2 ± 0.3) using chronic x-ray exposure as the reference radiation. The Johnson et al. [56] study primarily aimed to determine an RBE for tritium using myeloid leukemia as an endpoint and not necessarily as a measure of the induction of myeloid leukemia and other cancers. Therefore, while they reported on the incidence of other cancers, they did not determine the lowest dose of incidence. Other cancers such as reticulum cell sarcoma did occur, but usually at doses of 2 Gy or higher. Indeed, for some types of cancer, incidence appeared lower than in

controls in both the tritium and x-ray irradiated animals, but no tests for significance were done to verify this.

Seyama et al. [57] examined the cumulative incidence of tumors in mice over different exposure regimes for tritium beta irradiation, neutrons, Cobalt-60, and Cesium-137 gamma radiation. Mice were injected with HTO intraperitoneally at concentrations of 0.14, 0.28, 0.56, and 0.74 GBq, giving doses of 1.97, 3.95, 7.90, and 10.53 Gy, respectively. At 400 days, tumor incidence was 4%, 8%, 18%, and 24% respectively, but by 500 days, incidences differed little among the groups. This indicated that even the lowest dose (1.97 Gy) was sufficient to elicit the same incidence of tumors, with higher doses simply accelerating the onset. Increased incidence of T-cell lymphoma was also induced with 4 weekly injections of similar doses. Increased tumor incidence over controls was seen in the ovary, liver, lung, mammary glands, and uterus. There were also high incidences of lymphoma and leukemia.

Balonov et al. [58] summarized Russian studies that exposed mice and rats to tritium. One of the studies reported [59] exposed mice through continuous HTO ingestion of 37–1,850 kBq/g/day, which delivered doses of 1.2–2.8 Gy. Damage to DNA was only observed at the end of the intake period in animals with the greatest intake. DNA repair was also found to be slower in this group compared to the controls. The data presented by Balonov et al. [58] showed that tritium induced many types of tumors, including leukemia and solid cancers, in mice and rats.

A literature review by Straume [60] analyzed the risks, including cancer, of exposure to tritium radiation. Because cancer risk information due to tritium radiation was not available for humans, cancer risk estimates for tritium were derived from human populations exposed to gamma and x-ray radiation and from experimental animal studies. Straume [60] used a Monte Carlo sampling program to generate frequency distributions of excess cancer mortality from chronic, low-level exposure to HTO. He used frequency distributions for dose rate effectiveness factors from 1 to 12 (based upon animal studies), with a central value of the RBE for tritium between 2 and 3 and a range of 0–4.5. His analysis produced a skewed risk distribution with a 50th-percentile risk per

milligray of 81×10^{-6} with a 90% confidence range of $38\text{--}185 \times 10^{-6}$ per milligray. This value is comparable to radiation risk estimates of ICRP [61], BEIR [43], and UNSCEAR [62] which are 50×10^{-6} , 79×10^{-6} , and $70\text{--}110 \times 10^{-6}$ per milligray, respectively.

Although the above studies demonstrate that tritium's beta radiation is carcinogenic, it remains uncertain at what dose tritium will induce cancer. The lowest dose observed to induce cancer in mice is in the range of about 1 mGy/day. The work by Straume [60] perhaps gives the greatest insight into tritium-induced carcinogenesis and corroborates the risk factors presented by ICRP [61], BEIR [43], and UNSCEAR [62].

Hereditary Effects It is theoretically possible that damage to the chromosomes in oocytes or spermatogonia may be mutagenic and therefore carried forward in subsequent generations. However, the latest ICRP recommendations [63] maintain that "there continues to be no direct evidence that exposure of parents to radiation leads to excess heritable disease in offspring." Nevertheless, based upon animal experiment results, the ICRP estimates a genetic risk of about 0.2%/Gy for up to the second generation (grandchildren). For low-LET radiation, the ICRP value for the probability of severe hereditary effects is 0.5%/Gy, based on extrapolation from male mouse data.

The United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) [64] estimated genetic risks:

1. To a population sustaining low-LET, low-dose, or chronic irradiation exposure generation after generation
2. To a population that sustains low-LET, low-dose, or chronic irradiation to one generation only

In the first case, the total genetic detriment was estimated to be $\sim 0.41\text{--}0.64\%$ risk per parental Gy per million progeny above baseline (background incidence) to the first generation and $\sim 0.53\text{--}0.91\%$ risk per Gy per million progeny above baseline to the second generation.

In the second case, the total genetic detriment was again estimated to be $\sim 0.41\text{--}0.64\%$ risk per Gy per million progeny above baseline to the first generation. However, because only the first generation had received

the continuous radiation exposure, the risk in the second generation dropped to $0.16\text{--}0.43\%$ risk per Gy per million progeny above baseline.

As for the applicability of these risk estimates to tritium radiation, Straume and Carsten [25] noted that the genetic effects observed for other low-LET radiations were also present following exposure to tritium, either from HTO or from OBT. By grouping RBE studies with genetic endpoints (such as chromosome aberrations and mutations in mice), they determined that the RBE ranged from 1 to 3, with the higher values associated with low doses and low-dose rates (due largely to the curvilinear response of the reference radiation).

Therefore, it could be concluded that if gamma rays were the reference radiation, then the genetic risk estimates from UNSCEAR and the ICRP could be doubled for the risk posed by tritium exposure. It should be noted, however, that this conclusion is based upon animal studies and that genetic effects in humans due to radiation exposure have never been observed to date.

Adverse Health Effects of Organically Bound Tritium

Some of the tritium released into the environment will naturally bind with organic molecules, either through enzymatically catalyzed reactions where a hydrogen atom is replaced by a tritium atom or through natural exchange reactions. If the tritium atom is attached to a carbon atom in an enzymatically catalyzed reaction, it is essentially fixed to that molecule until the molecule is metabolized. Conversely, tritium atoms bonded to oxygen, sulfur, nitrogen, or phosphorous atoms will readily exchange with hydrogen atoms in the surrounding cellular water and are therefore considered as an exchangeable tritium compound [12]. Both nonexchangeable and exchangeable compounds are referred to as OBT. However, the nonexchangeable tritium will exhibit retention times determined by the metabolism of the molecules concerned, while the exchangeable tritium will be indistinguishable in its retention from that of HTO in cellular water.

OBT is a general term applying to a large group of compounds, where tritium is bound to organic molecules, including tritiated sugars, polysaccharides, lipids (fats), proteins, and DNA precursors. The extent of cellular injury caused by an OBT molecule depends largely upon where it is incorporated into a cell and if

it is there long enough for the tritium atom to decay. Tritiated DNA precursors, such as $^3\text{HTdR}$, are theoretically more efficient at causing cellular injury since they form part of the basic building block of a DNA strand. On the other hand, tritiated molecules that are more remote from DNA in the cell such as a tritiated fat molecule or a tritiated amino acid in a structural protein should pose a lesser risk.

Lambert [49] studied the effects of HTO and $^3\text{HTdR}$ on rat spermatogonia and determined an RBE of 1.3 for $^3\text{HTdR}$ and 2.3 for HTO. He pointed out that these values should be viewed with caution due to uncertainty on several factors, such as the time of death of the spermatogonia and the amount of tritium that induced it.

One of the few studies looking at effects of tritiated amino acids and nucleosides was done by Furuno-Fukushi et al. [65] where they treated mouse lymphocytic leukemia cells for 50 h with $^3\text{HTdR}$, lysine, arginine, leucine, and aspartic acid. Doses ranged from approximately 50 mGy to 800 Gy (as interpreted from graphical data). Cell survival decreased linearly for all compounds with effects greatest for thymidine, followed by arginine, lysine, leucine, and aspartic acid; similar results were found for cell mutation frequencies. The concentrations for detectable cell killing and mutagenesis ranged from less than 37 MBq/L for $^3\text{HTdR}$, 37–370 MBq/L for tritiated amino acids and 18.5–185 MBq/mL for HTO. Furuno-Fukushi et al. [65] compared the cell killing results of a previous study that examined the effects of HTO and gamma rays at similar doses and dose rates [66]; calculated RBEs of 2.9, 2.6, and 5.9 for mutation induction relative to low-dose rate gamma rays for HTO; tritiated amino acids and tritiated thymidine, respectively. However, the use of nonconcurrent reference radiation controls places some doubt on these results. They also reported that the tritiated amino acids were distributed homogeneously within the cell, as is the case for HTO.

Wang et al. [67] examined the effects of OBT on cultured mouse embryonic mid brain cells using $^3\text{HTdR}$, uridine, arginine, and glutamic acid. The cells were exposed to different concentrations of the tritiated compounds over a 20-hour period. Assays of cell proliferation and differentiation and DNA and protein content were conducted. To determine the RBE, cultures of cells were also exposed to 150 kVp x-rays at

0.5 Gy/min as the reference radiation. Calculated RBEs were in the range of 4.6–8.7 with the largest being for $^3\text{HTdR}$. The study was heavily criticized by Trivedi et al. [68] who pointed out various shortcomings; for example, the reference radiation was over a much shorter period than the tritium exposure, cell cultures were handled improperly, and the data analysis was missing information.

In their review of the available studies on OBT, Straume and Carsten [25] concluded that the RBE of tritium is slightly larger than that of HTO when it is bound to amino acids, and is about 2 times higher when bound to nuclear bases like thymidine.

In summary, while many studies have examined how OBT is partitioned within the body and within the cell, studies specifically looking at health effects due to OBT are limited. Those that are available indicate that most organic compounds have about the same RBE as HTO, since they are distributed throughout cells and do not lead to preferential irradiation of the nucleus. Incorporation of $^3\text{HTdR}$, however, can lead to concentration of tritium in the nucleus, resulting in RBE values of about twice that of HTO, calculated on the basis of average cell/tissue dose. With the exception of DNA precursors, it can be concluded that the greater effect of OBT relative to HTO is due to longer retention times, and proportionately greater dose per Bq intake, rather than differences in RBE (see sections “[Adverse Health Effects due to Tritium](#)” and “[Relative Biological Effectiveness](#)”).

Occupational and public doses due to exposures to OBT would arise from a mixture of tritiated compounds, including amino acids, sugars, fats, nucleic acids, and polysaccharides (in the case of public doses). There are some special instances where specific OBT compounds such as $^3\text{HTdR}$ would be used (for example, laboratory studies), but these would be under strictly controlled conditions to limit intakes. There are two relevant studies on this subject:

- Steel and Lamerton [69] measured tritium activity in several different organs in juvenile rats after an injection of $^3\text{HTdR}$. The greatest uptakes of $^3\text{HTdR}$ were in the bone marrow, the testis and the lungs. The lungs retained the $^3\text{HTdR}$ the longest, although at much lower concentrations than found in the bone marrow. The time for tritium activity to be

reduced to half the initial amount ranged from 1 day for the colon to 25 days for the testis, although the retention period changed over the course of the experiment (up to 56 days) likely due to the reuse of the $^3\text{HTdR}$.

- Lambert and Clifton [70] showed that only about 2% of ingested $^3\text{HTdR}$ is incorporated into DNA.

Human Epidemiology All of the studies mentioned previously have been under controlled, laboratory conditions and there are no significant studies of biota exposed to tritium under environmental settings. There have, however, been a number of human epidemiological studies that offer some insight on environmental exposures.

Epidemiological studies have formed the basis for attributing human health risks to radiation exposure. Studies of the Japanese atomic bomb survivors (Preston et al. [71–73]), patients treated with radiation (Boice et al. [74, 75]; Weiss et al. [76, 77]; Howe [78]; Howe and McLaughlin [79]; Little et al. [80]; Little and Boice [81]), and radiation exposed workers (Lubin et al. [82]; Cardis et al. [83]; Cardis et al. [84, 85]) have been assessed to estimate radiation risk coefficients (ICRP [61]; BEIR [24]; UNSCEAR [13, 51]. Unfortunately, there is little data specific to tritium. In epidemiological studies where tritium exposures exist, workers were also exposed to other radiations, particularly external gamma irradiation. Tritium specific exposures are often lacking from these assessments.

Several ecological studies of members of the public living in the vicinity of nuclear facilities assessed the relationship between the community's health (such as incidence and mortality rates of childhood leukemia, non-Hodgkin's lymphoma, birth defects, and other health indicators) and the average environmental releases from, or residential proximity to, a facility. There were no individual estimates of tritium exposure (or other risk factors) in these studies. No inference of the health effects from tritium can be made because of the limitations of such ecological studies [13]. Given the extremely low levels of radiation exposures from the facilities [86–91] adverse health effects would be highly unlikely.

The recent German KKiK case-control studies by Spix et al. [92] and Kaatsch et al. [93] of childhood

leukemia have raised significant attention. There were no such results in France [94] or Britain [95], and Grosche [96] observed that the trend in risk decreased over time, which could indicate some agent being involved, for which the prevalence is reduced over time. Grosche et al. [89] found no indication that tritium discharges from the Krümmel NPP were involved in any excess of childhood leukemia found nearby. Currently, there is no support for a causal relationship between any chemical or physical risk factor and the observed risk of childhood leukemia among children younger than 5 years of age living within 5 km of German NPPs (SSK [97]). The observed risk remains unexplained, however, it should be noted that clusters of childhood cancers, including leukemia, have been observed where no nuclear facilities exist.

The cohort studies of radiation workers provide the best available information on the risk of tritium exposure. However, few studies have direct information on tritium exposure, and fewer still directly assess the health effects of tritium exposure alone. The studies of Beral et al. [98], Cragle et al. [99], Johnson et al. [100], Zablotska et al. [101], Schubauer-Berrigan et al. [102], and Richardson and Wing [103] assessed health effects with tritium doses in combination with other (film-badge) doses. Thus, it is difficult to assess the component of risk associated specifically with tritium; especially since the tritium contribution to total dose was generally very small compared to other sources of radiation exposure. Hazelton et al. [104], for example, noted that whole-body tritium exposures were generally small in comparison to gamma exposures and that the dose-response for tritium in a restricted analysis was not statistically significant. The lack of an excess risk found from total occupational radiation dose, and the small contribution of tritium to the total dose, implies that any risk from tritium would be negligible [105].

The United Kingdom Committee Examining Radiation Risk from Internal Emitters (CERRIE) [106] report recognized an elevated health risk at moderate and high levels of internal exposure. However, at low levels any increase in risk may be undetectable and there is no clear evidence that current radiation risk estimates are radically underestimated. The AGIR [27] review of epidemiological studies of exposure to tritium concluded that the available epidemiological literature of environmental and occupational tritium exposures do

not provide enough detail to estimate risks from tritium exposure and recommended an international collaborative study of nuclear workers. The UK's Health Protection Agency [27] and the ICRP [107] supported the AGIR conclusions and recommendations.

In summary, while there has been a number of epidemiological studies conducted which have included tritium exposure, no definitive conclusions can be drawn as to effects to humans at current exposure levels. However, the lack of an excess risk from total occupational radiation dose and the small contribution from tritium to the total dose implies that any risk from tritium alone would be negligible.

Nonradiological Effects of Tritium

Transmutation Transmutation is the conversion of one element to another through radioactive decay. When tritium undergoes decay, it becomes Helium-3 (^3He), a stable, inert gas. As a different element, it no longer holds the chemical properties of a hydrogen isotope, which could lead to detrimental consequences in addition to those caused by the radiation. For example, if a tritium atom is bound to a DNA molecule when it decays, most of the kinetic energy will accompany the beta radiation as it is ejected from the nucleus, but some energy will provide a "kickback" to the ^3He atom as recoil energy. Kacena [108] determined that the recoil energy is too small (about 3 eV) to cause ionization of the DNA molecule on its own. However, the resultant ^3He atom would then be attached to a carbon atom with a very weak bond. This bond would tend to break leading to a free helium atom and possibly an ionized DNA molecule. The ionized DNA molecule may then be repaired, or if irreparable, lead to permanent DNA damage.

Myers and Johnson [109] performed a comprehensive review of transmutation effects. They pointed out that the degree of damage caused by transmutation of tritium to a helium atom could theoretically vary significantly, depending on the position of the tritium atom in specific DNA nucleotides. The studies also covered several test systems from the S13 virus, two strains of the bacterium *Escherichia coli*, *Drosophila melanogaster* (fruit fly), and cultured mammalian cells. Based on the position of the tritium on the nucleic acid, varying degrees of damage were noted – in one case, by a factor of 500 in the S13 virus. However, it should be

noted that the S13 virus only has a single strand of DNA and therefore does not have the repair capacity of higher organisms. In summary, Myers and Johnson [109] argued that it would be highly conservative to assume that the mutations observed in lower organisms would be detectable in mammalian cells, and that the number of detectable mutations in mammals (resulting from transmutation) would not likely exceed 5%.

Carsten [1] discussed the possibility that such effects would be manifested in humans after ingesting HTO or OBT as food. He suggested that the risk was small enough to pose no significant hazard, primarily because only 2% of the DNA hydrogen is located at the 5-position of the cytosine ring; therefore, damage would be minimal. In their review, Feinendegen and Bond [110] reached the same conclusion: "the effects of intracellular tritium are overwhelmingly due to beta irradiation of the nucleus" and "transmutation effects do not produce a measurable effect."

In any event, if DNA damage did occur from transmutation, it is unlikely that it could be discerned from radiation-induced damage and would be accounted for in the RBE. The AGIR [27] drew a similar conclusion.

Isotopic Effect of Tritium With two additional neutrons, a tritium atom has approximately three times the mass of a conventional hydrogen atom (protium). While chemically equivalent to hydrogen, tritium has slightly different physical properties due to its increased mass. Diabaté and Strack [12] noted that reaction rates decreased as atomic mass increased, causing a significant isotopic effect. Therefore, in reactions with HTO as a precursor, regular hydrogen would be preferentially used (instead of tritium) due to its lower atomic mass. However, as pointed out by the AGIR [27], any effects due to isotopic differences would be accounted for in the determination of RBE values.

Conclusion on Adverse Health Effects

Laboratory studies have demonstrated that tritium induces lethality, teratogenicity, genetic (chromosome aberrations), and reproductive effects. Tritium has also been shown to induce and promote cancer in mice under experimental conditions. Table 1 summarizes the studies reviewed in this report, although one should be careful when comparing the endpoints and the

Tritium, Health Effects and Dosimetry. Table 1 Overview of studies showing lowest concentration or administered dose of tritium to produce an adverse effect

Study	Means of exposure	Test animals	Parameter	Lowest concentration or administered dose to show effect	Dose (when available)
Lethality					
Brues et al. [17]	Single injection of HTO	Mice	LD _{50/30}	0.037 GBq/g body weight (BW) (~1.5 GBq total dose)	9 Gy (~6 Gy ^a)
Yokoro et al. [19]	Single injection of HTO	Mice	Lowest dose	0.37 GBq	~5 Gy
Yokoro et al. [19]	Single injection of HTO	Mice	LD _{50/30}	0.56–0.93 GBq	~8–13 Gy
Furchner [18]	Single injection of HTO	Mice	LD _{50/30}	0.024 GBq	8 Gy (~4 Gy ^a)
Yamamoto et al. [20]	Ingestion of HTO	Mice	Lowest dose	1.37 GBq	11.1 Gy
Teratogenic effects					
Zamenhof and Van Marthans [31]	Ingestion of HTO	Mice	No observed effects in mothers, temporary effects in offspring	111 MBq/L	N/A
Laskey et al. [32]	Ingestion of HTO	Mice	Lowest dose reduced brain weight	3.7 MBq/L	0.00003 Gy/day
Bursian et al. [33]	Ingestion of HTO	Mice	Reduced brain weight and neural hormones	370 MBq/L	0.66 Gy
Jones et al. [34]	Ingestion of HTO	Squirrel monkeys	Oocyte reduction, no other noted effects	111 MBq/L	N/A
Reproductive effects					
Satow et al. [44, 45]	Injection of HTO	Mice	Oocyte reduction	0.17 MBq/g BW	0.035 Gy
Lambert [49]	Injection of ³ HTdR	Mice	Reduction of spermatogonia	185 kBq/g BW	0.084 Gy (to cell nucleus)
Lambert [49]	Injection of HTO	Mice	Reduction of spermatogonia	2.22 MBq/g BW	0.049 Gy (to cell nucleus)
Carcinogenicity					
Yamamoto et al. [53]	Ingestion of HTO	Mice	Thymic lymphoma	1.85 GBq/L	0.048 Gy/day
Yamamoto et al. [53]	Ingestion of HTO	Mice	Solid tumors	0.925 GBq/L	0.024 Gy/day
Yamamoto et al. [54]	Ingestion of HTO	Mice	Thymic lymphoma	0.463 GBq/L	0.012 Gy/day
Gragtmans et al. [55]	Injection of HTO	Rats	Early onset of mammary tumor	0.45 MBq/g BW	0.46 Gy
Johnson et al. [56]	Injection of HTO	Mice	Myeloid leukemia	90 MBq	0.85 Gy
Seyama et al. [57]	Injection of HTO	Mice	Tumors	140 MBq	1.97 Gy
Mushkacheva et al. [59]	Ingestion of HTO	Mice	DNA damage	37 kBq/g/day	1.2 Gy

^aEstimated for this report using a biological half life of water of 1.2 days

different dose regimes (injection or ingestion), especially with respect to the mouse oocyte. Generally, doses above 0.5 GBq and radiation doses of 0.5 Gy are needed to induce an adverse health effect. The exception to this, as discussed earlier, is the mouse reproductive system, whose oocytes are about 100 times more radiosensitive than that of humans.

Because it is ionizing radiation, the tritium beta particle would also be expected to induce hereditary and reproductive effects. This is in line with an equivalent dose of gamma or x-ray radiation, although no studies were found to support this hypothesis. Similarly, noncancer effects such as cardiovascular disease could also be expected as a result of radiation exposures to doses exceeding 0.5 Gy. Not surprisingly, the doses resulting from the amount of tritium needed to cause these effects are essentially the same as doses of other types of ionizing radiation (gamma and x-ray) when the RBE and dose protraction are taken into account. Current risk estimates for tritium therefore appear to be appropriate.

In general, the greater biological effectiveness of OBT relative to HTO appears to be related to longer retention times, rather than to differences in RBE. DNA precursors are the exception, where preferential irradiation of the nucleus can give higher RBE values, expressed in terms of average cell dose. However, OBT concentrations in the environment are generally low and the compounds that pose the greatest risk (such as $^3\text{HTdR}$) are not expected to approach concentrations anywhere near those that pose a significant risk.

Nonradiological effects from tritium such as transmutation and isotopic effects are unlikely, but if they do occur, it would not be possible to distinguish their effects from those of radiation. Therefore, they are taken into account in risk estimates and do not require special consideration.

Relative Biological Effectiveness

In radiation biology, absorbed dose refers to the amount of radiation energy deposited within a unit mass, typically tissue or individual cells. However, it is clear that the deposited energy alone does not dictate the amount of biological damage and that other factors like the type of radiation and the chemical form of the radioisotope

play a role. To equate absorbed doses from different radiations, and thus permit the use of a single unit of dose for radiation protection purposes, the effectiveness of the radiations must be quantified. Quantification requires a strict comparison of the effects of the radiation of interest with those effects caused by a standard or reference radiation. The ratio of the response of the two radiation types is termed the RBE.

With respect to tritium beta radiation, the selection of the reference radiation makes a significant difference in the resultant RBE. In addition, Little and Lambert [111] and the AGIR [27] point out that the dose from the reference radiation must be delivered in the same fashion as that of the tritium irradiation, which implies a chronic exposure with an exponential decay if the concentration of the tritium in the body is not maintained at a constant value.

To further complicate matters, the response of the tissue or organ being irradiated may not be linear for the radiation of interest or the reference radiation. Straume and Carsten [25] discuss this at some length, indicating that the RBE can vary significantly with changes in dose rate and radiation quality (LET). Therefore, for radiation protection purposes, it is desirable to assess the RBE for doses and dose rates that are comparable to those received in environmental and occupational settings. However, as the AGIR [27] points out, the dose response curve for acute doses of low-LET reference radiation is often curvilinear in relation to the response of radiations with higher LET (such as tritium's beta radiation). Studies at these lower doses give a maximum RBE value, which would be more appropriate for occupational and environmental exposures. This is referred to as the RBE_{MAX} or RBE_{M} .

There have been many experimental determinations of the RBE for tritium radiation, using various biological endpoints, both in vitro and in vivo, and using gamma and x-rays as the reference radiations. In vivo (Latin for "within the living") implies that the living tissue of an entire living animal is used in the study. In vitro (Latin for "within the glass") is the general designation given to experimental studies done on tissues or cells outside of the body; for example, in a Petri dish.

Several reviews of these studies have been conducted, some of them critical, that have highlighted

errors or omissions in the work that might have affected the RBE determination. These too are briefly reviewed in the following section.

RBE Literature Reviews and Experimental Studies

Straume and Carsten [25] identified 12 published RBE studies for HTO compared with orthovoltage (200–500 kVp) x-rays and 21 RBE studies for HTO compared with gamma rays from Cesium-137 or Cobalt-60. Combining these studies, they calculated an arithmetic mean RBE value of 1.8 using x-rays as the reference radiation and 2.3 relative to gamma rays. They commented that their estimates were based upon studies that used radiation delivered at doses and dose rates higher than those generally received by radiation workers and members of the public; therefore, their estimated RBE values could have been lower than the actual RBE values than should be used for human risk assessment.

The Little and Lambert [111] critique of the Straume and Carsten [25] review pointed out that Straume and Carsten [25] included a number of studies that did not have adequate concurrent external (x-ray or gamma ray) radiation-exposed groups (e.g., Prosser et al. [112]; Carsten and Commerford [113]; Russell et al. [114]; and Furuno-Fukushi et al. [65]). In addition, some of the other calculations of RBE conducted by Straume and Carsten [25] (for example, the tritium experiments of Prosser et al. [112]) combined with the Cobalt-60 experiments of Lloyd et al. [115] were based on nonconcurrent experiments.

The AGIR report [27] discusses the RBE of tritium radiation as well as track-structure considerations, the relationship between the RBE and the dose and dose rate effectiveness factor (DDREF), the use of different reference radiations, and experimental studies. The biokinetic models of HTO and OBT were discussed through an examination of the current ICRP models, a review of the key information underlying those models, a discussion of more recent models, and the special aspects of DNA precursors. There was also a detailed review of reproductive effects in the female, especially on oocyte formation and fertilization.

The AGIR authors critically reviewed several papers that estimated the tritium RBE value under different experimental regimes. Twenty-four studies were

divided into *in vivo* and *in vitro* experiments, and subdivided as either chronic or acute irradiation regimes. To be classified as a chronic irradiation, the external irradiation dose rate had to be comparable to the tritium exposure. The biological endpoints of the tritium studies included carcinogenesis, chromosomal aberration, cell death, and others. With one exception, they only included papers published in peer reviewed journals or conference proceedings; the exception being an article by Chopra and Heddle [116], which they included because of the well-conducted experiments and data analysis.

The AGIR made the following conclusions regarding the tritium RBE value:

1. Various theoretical and experimental studies of radiation of LET similar to that of tritium beta particles led to the general expectation of an RBE value for tritium of at least 2 compared with gamma radiation.
2. Neither transmutation effects nor isotopic discrimination associated with tritium appear likely to have major effects, but any effect would be in the direction of tending to increase the observed RBE value.
3. Experimentally determined RBE values can vary significantly depending on the choice of reference radiation. It was recommended that high-energy gamma rays should be the preferred choice for reporting RBE values. Where lower energy x-rays and gamma rays were used as the standard, an appropriate adjustment should be taken into account when discussing the results. In addition, the reference radiation used should be adequately described and include anything, such as filtration, that would modify the energy spectrum.
4. Interpretation of RBE experiments was complicated by dose rates that were rarely comparable and by reference radiation that may have been more effective than hard gamma rays.
5. In a wide variety of cellular and genetic studies, RBE values for HTO have generally been observed in the range from 1 to 2 when compared with orthovoltage x-rays, and in the range from 2 to 3 when compared with gamma rays.
6. For developmental endpoints, the RBE values for HTO were similar to those obtained from cellular and genetic studies.

7. Whole animal carcinogenesis studies yielded RBE values with central estimates generally in the range of 0.8–2.5 (x-rays and gamma rays as the reference radiations). However, the AGIR had several reservations with these studies. In particular, in some of the studies the frequency of cancers appears to have been saturated or nearly saturated at the lowest doses employed (about 1 Sv), so the slope of the dose–response curve was above the relevant dose levels – which represent the area of interest for radiation protection purposes.
8. The AGIR considered the most likely RBE value relative to chronically delivered hard gamma rays was between 2 and 3 and considered a value of 2 most appropriate. This was based largely on an analysis of the available experimental data with rounding and biophysical considerations. Fractional values were not considered appropriate.

In addition to the above, the most significant recommendation of the AGIR report was related to the use of a provisional value of 2 for the radiation weighting factor (w_R) for tritium until an internationally agreed value becomes available.

Little and Lambert [111] used the RBE estimates from the acceptable studies to compute an aggregate “best linear unbiased estimate (inverse-variance weighted) of the RBE” and the 95% confidence limit of the RBE estimates. Their definition of the RBE was based upon the ratio of the initial slopes (i.e., lower doses) of the tritium and gamma (or x-ray) dose–response curves which was roughly equivalent to the standard definition of the RBE_M .

In their summary of the carcinogenesis studies, Little and Lambert [111] indicated that the overall aggregate results imply RBE values with a central estimate in the range of 1.2–2.5 and 97.5 upper percentile of no more than 3.0. The three studies that used chronic x-rays and chronic gamma irradiation yielded an aggregate RBE value of 1.19 (95% confidence interval (CI): 0.88–1.49) and 2.49 (95% CI: 2.00–2.98), respectively.

Little and Lambert [111] stated that they doubted the in vivo cancer data in the studies. Nevertheless, their estimates of RBE in relation to late health effects (such as cancer) were derived by combining the in vivo cancer data estimates for endpoints for other than cell

survival (and related endpoints) to yield an aggregate RBE of 2.19 (95% CI: 2.04–2.33) and 1.17 (95% CI: 0.96–1.39) when using chronic gamma rays and chronic x-rays, respectively, as the reference radiation.

The term “radiation effectiveness factor” (REF) was used by Kocher et al. [117] to represent the biological effectiveness of a radiation dose for the purpose of estimating cancer risks and probability of causation in identified individuals. This differs from an RBE in that the REF applies strictly to the results of specific radiobiological studies under controlled conditions. Also, the REFs for induction of cancers in humans developed in this work were expressed as both a central estimate (median) and a subjective probability distribution (95% confidence levels) to represent their uncertainty.

Part of the work by Kocher et al. [117] estimated the RBE_M for tritium beta particles, where M represents the maximum value. Estimates of the RBE_M were taken from Tables 1–3 of Straume and Carsten [25] where three of five studies, using chronic exposure to x-rays, had a central estimate of 2–3 for RBE_M . They further reasoned that “if a nominal biological effectiveness of X-rays relative to high-energy gamma rays of 2 is assumed on the basis of the probability distribution of REF_L for orthovoltage X-rays . . . , those data suggest that RBE_M for 3H beta particles relative to gamma rays delivered chronically could be as high as about 4–6.” Based on this reasoning, their REF_L for “electrons less than 15 keV had a 50th percentile value of 2.4 and a 95% confidence interval of 1.2–5.0.”

The *Canadian Environmental Protection Act*, 1999 (Environment Canada [118]) requires the federal Ministers of the Environment and of Health to prepare and publish a Priority Substances List (PSL) that identifies

Tritium, Health Effects and Dosimetry. Table 2 Track average LET in water for various radiations based on a cut-off energy, Δ , of 100 eV (based on ICRU [119])

Radiation	$\bar{L}_{\Delta,T}(\text{keV}/\mu\text{m})$
Cobalt-60 gamma radiation (1,173 and 1,332 keV)	0.22
200 kVp x-rays	1.7
Tritium beta radiation (mean 5.7 keV)	4.7
50 kVp x-rays	6.3

Tritium, Health Effects and Dosimetry. Table 3 ICRP Publication 67 – [127]-Parameters describing the distribution and retention of tritium after acute intakes of tritiated water

Age	Distribution (%)		Biological half-life (days)	
	HTO component	OBT component	HTO component	OBT component
3 months	97	3	3.0	8
1 year	97	3	3.5	15
5 years	97	3	4.6	19
10 years	97	3	5.7	26
15 years	97	3	7.9	32
Adult	97	3	10.0	40

substances, including chemicals, groups of chemicals, effluents, and wastes, that may be harmful to the environment or constitute a danger to human health. In complying with the Act, the Ministries of Environment Canada and Health Canada undertook an assessment of the impact on nonhuman biota from the releases of radionuclides from nuclear facilities (Environment Canada/Health Canada [118]). Tritium was included in the review since it is released to the environment from the operation of CANDU nuclear power reactors. Although nonhuman biota were the primary focus, much of the review included studies looking at the effects of tritium on mammals and so is relevant to this report. Reproductive effects of tritium exposure were of special interest, since reproductive organs are one of the most radiosensitive tissues and reproductive failure could affect biota populations. The report looked at many studies and reviews to determine what the investigators called an “ecodosimetry weighting factor.” They reported the results of nine studies looking at reproductive effects of tritium in mammals and fish, which gave tritium radiation RBE values from 1.7 to 3.8 (the highest value was due to reduction in fish egg fertilization). The report also “gamma normalized” (multiplied the RBE by 2, using x-rays as a reference to make the value comparable to a gamma radiation reference) the RBE value of spermhead survival reported by Lambert [49], giving an estimated RBE of 4.8. These RBE values were compared to those of Straume and Carsten [25], resulting in an “ecodosimetry weighting factor” of 3 for tritium radiation for ecotoxicological risk assessment purposes.

Electron Energy and Ionization Path Lengths: Why Tritium Radiation Is More Effective Than Gamma Radiation The experimental evidence seen in this and other reviews provides a convincing argument that the RBE for tritium radiation is greater than that of 200–250 kVp x-rays and Cobalt-60 or Cesium-137 gamma radiation.

The amount of energy lost per unit distance is referred to as the linear energy transfer (LET), often expressed in units of keV/μm. Interactions between radiation and molecules will cause ionization in some of the molecules. Electromagnetic radiation, such as gamma rays and x-rays has a much lower LET than alpha particles. As Table 2 indicates (International Commission on Radiological Units and Measurements ICRU [119]), low-energy photons and electrons release more energy per micrometer than higher energy photons. However, the lower energy photons release much less total energy and travel much shorter distances than their higher energy counterparts.

Photons (gamma rays and x-rays) act indirectly and the type of interaction is energy related, causing molecular ionization through photon absorption (photoelectric and pair production) or scattering events (Compton interactions). Particle radiation, such as tritium’s beta radiation, can damage key molecules like DNA through both direct and indirect interaction. This damage occurs from the production of free radicals (e.g., hydroxyl) caused by the ionization of water molecules – which may, in turn, damage DNA.

The most significant cellular damage caused by radiation is breakage of the DNA molecular chain. DNA is a double helix, meaning there are two

corresponding and linked molecular chains containing corresponding nucleic acids. Radiation-induced DNA damage usually results in damage to the base molecules; single strand breaks (SSBs), double strand breaks (DSBs), DNA-protein cross-links, or a combination of some or all of these. The degree of damage is related to the density of the radiation track. A large deposition of energy within a DNA molecule may cause a cluster of damage. If the damage is great, then it may not be fully repairable and can cause cell death or permanent chromosome damage.

Nikjoo and Goodhead [120] suggested that the most significant biological effects are caused by sizeable clusters of energy deposition and that the net biological effect increases with the severity of the initial damage. Most of this type of damage is thought to be due to low energy secondary electrons that are produced as the radiation slows down. They concluded by stating that the lower energy electrons are more effective than higher energy ones (for example, greater than 100 keV) in producing complex damage and that this was observed in experiments with ultrasoft x-rays.

Another important aspect is the location of the tritium atom when it disintegrates. HTO will be relatively evenly distributed throughout tissues and cells. On the other hand, OBT may be incorporated into important molecules such as DNA. The release of a beta particle from that position has a greater probability of causing damage than an HTO molecule in the cellular fluid, for example. With respect to the impact of OBT, Chen [120] used a Monte Carlo code to generate simulated tracks in order to compare the dose mean linear energies (keV/ μm) for HTO and OBT molecules in spherical regions with diameters of 10 nm to 2 μm . The results indicated that a mean linear energy for OBT was a factor of 1.7 higher than for HTO, using the assumption that the extent of the increase would depend on the OBT molecule's location within the cell.

Computational models appear to confirm that the higher RBE of tritium, as well as that of other low-energy radiation, is at least partly due to the high-LET and correspondingly short path length of tritium beta particles. The relatively large deposition of energy within a very small volume induces SSBs and DSBs and also causes complex damage that may not be repairable. When the tritium is incorporated into DNA, there is a higher probability of biological damage.

Summary

The RBE provides a referenced correction factor for comparing the effects of different radiations, giving direction for the development of a universal measurement of dose, the sievert (Sv). Much of the discussion in this report and those of others, such as the AGIR [27] and Straume and Carsten [25], has centered upon an appropriate single value for the RBE of tritium. Although there have been more than 50 different estimates of this value, it is clear considerable variation exists. This, combined with uncertainty in radiobiological data, makes it difficult to select one value. The choice of a reference radiation accounts for much of this variability, due to the difference in the RBE of x-rays and higher energy gamma rays. Other factors contributing to the observed range include:

- Differences in radiosensitivity of tissues, organs, and organisms
- Different biological endpoints
- Variation in DDREFs
- Dosimetry
- Choice of in vitro or in vivo test systems

For occupational and public exposures, the most relevant endpoint is radiogenic cancer, (particularly radiation delivered at occupational or slightly greater levels), but there are very few studies at occupational dose levels, since most experiments use more than 1 Gy. However, overwhelming evidence supports an RBE for tritium of at least 2 when pooling the data from all studies, including those examining the most relevant endpoints, which were reviewed in this report. Given the range of observed values, an RBE of at least 2 for tritium beta particles should be seen as a general indication of the greater effectiveness of these particles versus low-LET radiation. This would apply to cancer induction in humans in the absence of better experimental data and specifically, any tritium-specific epidemiological data.

Dosimetry and Biokinetics

Calculating the committed effective dose received as a result of taking tritium into the body is done in two steps. First, a model describing the behavior of the tritiated compound taken into the body is used to predict where the tritium will go in the body and the rate at

which it will be removed from the body. Such a model is known as a biokinetic model. Second, a dosimetric model is used to calculate how much tritium will decay while it is in the various organs and tissues of the body and how much of the energy from the beta particles it emits during decay will be absorbed by the body. This allows us to calculate the dose resulting from the intake of tritium. Because tritium is removed gradually from the body after an intake, the committed effective dose takes into account the dose received for 50 years following the intake of a radionuclide into the body.

Since the energy carried by tritium beta radiation is of short range, all of its energy is absorbed in the tissues and organs in which the tritium atoms are located. This describes the dosimetric model for tritium. The biokinetic model for tritium is more complex and depends on the type of tritiated compound taken into the body.

This section presents the main biokinetic models used for calculating dose as a result of taking tritium compounds into the body, either by inhalation, ingestion, or absorption through the skin. In particular, it presents the tritium biokinetic models recommended by the International Commission on Radiological Protection (ICRP) and the basis for these models. Other models are also considered, which take into account types of exposures that are not considered by the ICRP models or give a more accurate description of the biokinetics of certain tritiated compounds. Examples of applications of the tritium dosimetric model are also provided by presenting dose coefficients calculated with the biokinetic models discussed.

Overview of Current ICRP Models for Tritiated Water and Organically Bound Tritium

Tritium may enter the body by inhalation, absorption through skin, and ingestion. The first two are the most frequent forms of intake in the workplace, while all three routes of intake can contribute to exposures to members of the public. Skin contact with tritium-contaminated surfaces, such as metal and glass, has been shown to result in the formation of OBT in the body. Skin contact is therefore considered as a different mode of intake compared to the absorption of HTO through skin, and has also been shown to be a route of intake in the workplace [7].

The fate of tritium once taken into the body is determined mostly by its chemical form in the external environment. One can expect to find HTO in most workplaces and environmental media where tritium is present. In most OBT compounds, tritium that is bound to oxygen, nitrogen, phosphorus, or sulfur will exchange freely with hydrogen in water. Therefore, it has the same metabolism and distribution in the body as HTO and is called the exchangeable bound tritium fraction. Tritium bound to carbon will not exchange with hydrogen in water. Tritium in such C–H bonds is known as nonexchangeable bound tritium. In general, considerations of the biokinetics of OBT refer to the nonexchangeable component that exhibits retention times determined by carbon turnover rather than HTO kinetics.

Tritiated compounds may also exist in particulate forms, such as airborne particles that contain tritium. The retention and clearance of these particulates in the respiratory tract depend on several factors, for example, the particle's size and chemical composition. For the purpose of dose assessment, tritium absorbed into blood after tritiated particulates have been inhaled is treated as HTO after it has left the lungs [122–125]. Although tritium beta rays are partially self-attenuated within the particle, the dose to the lower lung – from particles that slowly dissolve – can be up to two orders of magnitude higher than that from the same activity of inhaled HTO [126].

The tritiated compounds are therefore categorized according to the metabolic model that best describes their dynamics after uptake. Two primary metabolic models, which are discussed in this section, are used to estimate the dose from tritiated compounds:

- *The HTO model*, which is used to estimate the dose resulting from intakes of HTO or other tritiated compounds that partially convert to HTO after being taken into the body
- *The OBT model*, used to estimate the dose resulting from intakes of various tritiated organic compounds (nonexchangeable bound tritium)

A third model, which applies to tritium absorbed through the skin from HT-contaminated surfaces, is also described. Finally, the biokinetics of tritiated compounds in relation to pregnancy and to nursing is discussed in this section.

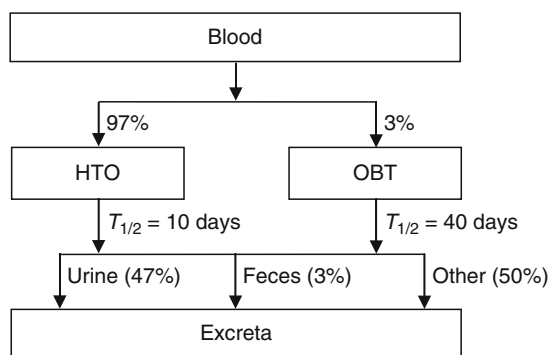
Internal Dosimetry of Tritiated Compounds Using the ICRP HTO Model

The ICRP Model for the Inhalation and Ingestion of Tritiated Water The ICRP HTO model [127], illustrated in Fig. 2, is used to assess doses for:

- Intakes of HTO
- HTO formed following intakes as HT
- Tritiated hydrocarbon vapors and gases
- Tritiated particulates

For intakes as HTO, the model assumes instantaneous translocation to blood. It is further assumed that HTO is transferred from the blood, with a biological half-life of 6 h and then distributed uniformly throughout the body. The model assumes that 97% of tritium taken in remains as HTO once distributed, while 3% is converted to OBT. In adults, HTO is retained with a biological half-life of 10 days, and OBT is retained with the biological half-life of carbon, which is 40 days. HTO is assumed to be distributed uniformly throughout the 42 L of body water. The assumed partitioning between HTO and OBT after acute intakes of HTO in various age groups, as well as corresponding biological half-lives, is given in Table 3 [128].

The model used to derive the committed effective dose per unit intake (dose coefficient) to adults resulting from the intake of HTO, as recommended by the ICRP [125, 128] and is based on Fig. 2. This model considers the ICRP's recommendations for radiation weighting factors as well as tissue weighting factors. The committed effective dose per unit intake is



Tritium, Health Effects and Dosimetry. Figure 2 ICRP Publication 71 – [125]–Model for the biokinetics of tritiated water

the computed effective dose received up to 50 years following a single intake for adults and up to 70 years for intakes by infants and children. The value for intakes of HTO by adults computed by the ICRP is 1.8×10^{-11} Sv/Bq. A slightly different dose coefficient, 2.0×10^{-11} Sv/Bq, was recommended by Health Canada and the Canadian Nuclear Safety Commission [129, 130] and incorporated in past regulatory documents [131]. The ICRP dose coefficient is based on the assumption that dose from HTO is received by 68.8 kg of tissue, which includes several kilograms of noncellular bone. This value was derived from the total body mass of 70 kg minus the masses of the contents of the gastrointestinal tract (1,005 g), urinary bladder (102 g), and of the gall bladder (62 g) [5]. Because the relevant dose is that to the nucleus of sensitive cells from contained HTO, the Atomic Energy Control Board (now the Canadian Nuclear Safety Commission) recommended using a dose coefficient of 2.0×10^{-11} Sv/Bq to assess the dose resulting from intakes of HTO. This recommendation considered dose received by 63 kg of soft tissue: a mass equal to the mass of the total body (70 kg), minus the skeleton mass (10 kg). Because the latter includes radiosensitive marrow, the mass of the marrow (3 kg) must be included in the dose calculation. The dose coefficient of 2.0×10^{-11} Sv/Bq considered dose to a mass of $70 \text{ kg} - 10 \text{ kg} + 3 \text{ kg} = 63 \text{ kg}$ [5]. Furthermore, the dose coefficient recommended by the Federal Provincial Working Group on Bioassay and In Vivo Monitoring [129], 2.0×10^{-11} Sv/Bq, assumed that 97.8% of tritium entering the body as HTO remained as HTO, with a biological half-life of 9.7 days, while 2.2% was converted to OBT and was retained with a biological half-life of 48.5 days [132].

More recently, the ICRP model for HTO [128] was used to calculate the dose to 68.8 kg of soft tissue, with corrections using the updated values of 10.5 kg for skeleton mass and 3.65 kg for marrow mass [133]. The resulting dose coefficient was 2.0×10^{-11} Sv/Bq, a value that the CNSC continues to recommend and use [130].

The contribution of the OBT fraction to dose in the ICRP's HTO model has been shown to be about 10% [132, 134]. This has been verified to be adequate for dose estimates under both acute and chronic intakes [135, 136].

Absorption of HTO Through Skin HTO is also known to be absorbed through skin from the vapor phase or the liquid phase, i.e., by immersion [3, 4, 137]. The ICRP [125] addresses HTO intake via skin absorption by reference to Osborne [137], Hill and Johnson [7], and Myers and Johnson [109], the latter being a literature review. About 1% of HTO activity per m³ in air is taken to be absorbed through the skin per minute. This results in absorption through the skin contributing about one third of the total HTO intake for a given HTO concentration in air when the exposed individual is active during exposure (breathing in more air than when at rest). Using the same assumptions, the amount of HTO absorbed through the skin in a given time is about equal to the amount inhaled in the same period of time when the exposed individual is at rest during exposure.

ICRP Model for the Inhalation of Elemental Tritium Gas Following inhalation of elemental tritium gas (HT), a small fraction (about 0.01%) [138] of the inhaled activity is dissolved in body fluids and then oxidized to HTO. The latter is the predominant contributor to dose. HT gas is not significantly absorbed through the skin and does not readily convert to HTO on the skin. Irradiation of the lungs from inhaled HT gas does not significantly increase the committed effective dose [139] because of the short range in tissue of this tritium beta particle (see Section The Inhalation of Elemental Tritium Gas). The dose coefficient for the

inhalation of HT gas is therefore taken to be 0.01% of that used for the inhalation of HTO.

Tritiated Hydrocarbons Various forms of tritiated hydrocarbons have been identified. Tritiated methane (CTH₃) is the only such tritiated compound for which the ICRP recommends a dose coefficient based on the HTO model. CTH₃ is known to form as a result of microbial degradation within tritiated waste. About 1% of inhaled CTH₃ is assumed to be converted to HTO [140]. The dose coefficient for CTH₃ is therefore 1% of that for HTO. The HTO model is not recommended for interpreting bioassay data collected following an intake of CTH₃; the use of workplace or environmental data is more appropriate. The ICRP approach is considered to give a conservative dose coefficient [140].

Summary of Dose Coefficients Based on the ICRP HTO Model Table 4 presents the ICRP's recommended dose coefficients for tritiated particulates, as well as for other tritiated compounds [141].

Table 5 illustrates the effect of age on the dose per unit intake for the inhalation of HTO and compares these dose coefficients to the value for adults.

Internal Dosimetry of Tritiated Compounds Using the ICRP OBT Model

Figure 3 illustrates the ICRP model [125, 127] used to assess doses that result from the intake of unspecified

Tritium, Health Effects and Dosimetry. Table 4 ICRP Publication 72. [141]. Dose coefficients based on the ICRP HTO model for various tritiated compounds, modes of intake and age groups

Tritiated compound	Mode of intake	Dose coefficient (Sv/Bq)	
		Infants (1 year old)	Adults
HTO ^a	Inhalation	4.8×10^{-11}	1.8×10^{-11}
HTO	Ingestion	4.8×10^{-11}	1.8×10^{-11}
HT	Inhalation	4.8×10^{-11}	1.8×10^{-15}
CTH ₃	Inhalation	4.8×10^{-13}	1.8×10^{-13}
Type F particulates ^b	Inhalation	2.0×10^{-11}	6.2×10^{-12}
Type M particulates ^b	Inhalation	2.7×10^{-10}	4.5×10^{-11}
Type S particulates ^b	Inhalation	1.0×10^{-9}	2.6×10^{-10}

^aThe dose from the inhalation of HTO does not include absorption through the skin and should be increased by a factor of 1.5 in order to account for it
^bDose coefficients shown for particulates of 1 μm activity median aerodynamic diameter (AMAD)

organic forms of OBT. It assumes that OBT, once taken into the body, is translocated to blood completely and instantaneously and then distributed uniformly throughout body tissues. In the absence of specific information on the inhaled organic compound, it is assumed that tritium absorbed to blood will behave as it does after ingestion (i.e., that 50% is retained as OBT in body tissues and that 50% is catabolized to HTO) and that its turnover time will be similar to those of water.

It is assumed that tritium inhaled as organic compounds is absorbed to blood instantaneously and is transferred from blood with a biological half-life of 6 h, with 50% retained as OBT in body tissues and 50% catabolized to HTO [142].

Data on the metabolic processes and retention of nonbiological tritiated organic compounds (such as

tritiated oils and solvents) following inhalation are scarce [7]. These compounds are therefore not considered further in this section.

Tritiated organic compounds in the diet may include tritiated proteins, fats, and carbohydrates. The ICRP model for the ingestion of OBT [127] is intended to represent normal dietary content of these different forms. The model was developed in the absence of good information about the specific proportions of OBT in components of the human diet and how tritium is incorporated into organic molecules in body tissues.

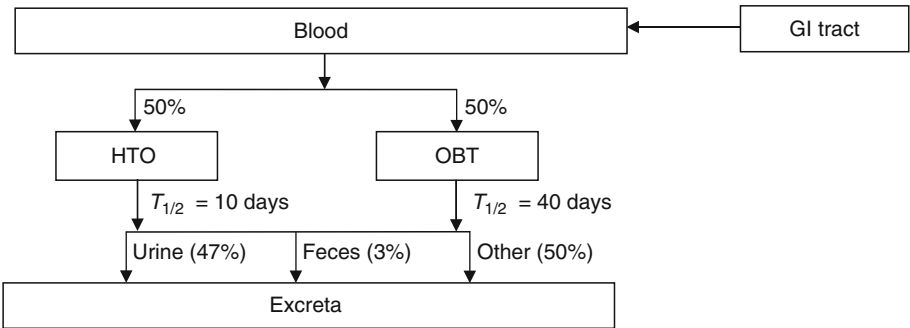
In this model, it is assumed that all ingested tritiated organic compounds are completely absorbed from the gastrointestinal tract to blood, after which the tritium from these compounds is uniformly distributed in tissue. In reality, the distribution of tritium in the various tissues will depend on the chemical form of the tritium and the metabolic activity of the individual tissue.

Experimental studies of the relative incorporation of tritium into OBT in body tissues following intakes of HTO and OBT suggest that about 9–45% of tritium from dietary OBT is incorporated into body OBT. The ICRP OBT model therefore conservatively assumes that after tritium is absorbed from blood (in adults), further to ingestion as OBT:

- Fifty percent of tritium will follow the metabolic behavior of carbon, having a biological half-life of 40 days.
- Fifty percent of tritium will behave as HTO, having a biological half-life of 10 days.

Tritium, Health Effects and Dosimetry. Table 5 ICRP Publication 72 [141]. Dose coefficients for HTO inhalation for various age groups

Age	Dose coefficient (Sv/Bq)	Ratio of dose coefficient to that of adult
3 months	6.4×10^{-11}	3.6
1 year	4.8×10^{-11}	2.7
5 years	3.1×10^{-11}	1.7
10 years	2.3×10^{-11}	1.3
15 years	1.8×10^{-11}	1.0
Adult	1.8×10^{-11}	1.0



Tritium, Health Effects and Dosimetry. Figure 3 ICRP Publication 71 – [124]-Model for the biokinetics of organically bound tritium

Tritium, Health Effects and Dosimetry. Table 6 ICRP Publication 67 [128]. Partitioning between HTO and OBT after dietary intakes of OBT

Age	Distribution (%)		Biological half-life (days)	
	HTO component	OBT component	HTO component	OBT component
3 months	50	50	3.0	8
1 year	50	50	3.5	15
5 years	50	50	4.6	19
10 years	50	50	5.7	26
15 years	50	50	7.9	32
Adult	50	50	10.0	40

It should be noted that less than 10% of OBT in diet is excreted in the form of OBT in urine (mostly in the form of urea), and ~3% OBT is excreted in feces.

Table 6 gives the assumed partitioning between HTO and OBT and corresponding biological half-lives, after dietary intakes of OBT in various age groups [128]. The ICRP's recommended dose coefficients for OBT are shown in Table 7 [141]. Table 8 shows the effect of age on dose per unit intake for the ingestion of OBT.

Intakes of Tritium in Relation to Pregnancy and Nursing

Pregnancy and Tritium Intake The ICRP [142] has provided biokinetic and dosimetric models and dose coefficients for the embryo, the fetus, and the newborn as a result of intakes of radionuclides by the mother.

In prenatal dosimetry, the term "embryo" refers to the developing offspring up to the end of week 8 of pregnancy, including the initial stages of growth up to the end of organogenesis. At this time, the embryo weighs less than about 10 g. During this stage, the dose to the fetus is assumed to be the same as that to the uterus wall.

The term "fetus" refers to the developing offspring after week 8 of pregnancy. During this stage of pregnancy, doses to fetal organs and tissues are calculated using biokinetic models that describe tritium behavior in the fetus itself, as this behavior will differ from the assumed tritium behavior in the mother. For the ICRP [142] report, models that take into account existing relevant biokinetic data were developed to calculate

Tritium, Health Effects and Dosimetry. Table 7 ICRP Publication 72 [141]. Dose coefficients based on the ICRP OBT model for various modes of intake and age groups

Tritiated compound	Mode of intake	Dose coefficient (Sv/Bq)	
		Infants (1 year old)	Adults
OBT	Inhalation	1.1×10^{-10}	4.1×10^{-11}
OBT	Ingestion	1.2×10^{-10}	4.2×10^{-11}

Tritium, Health Effects and Dosimetry. Table 8 ICRP Publication 72 [141]. Dose coefficients for OBT ingestion for various age groups

Age	Dose coefficient (Sv/Bq)	Ratio of dose coefficient to that of adult
3 months	1.2×10^{-10}	2.9
1 year	1.2×10^{-10}	2.9
5 years	7.3×10^{-11}	1.7
10 years	5.7×10^{-11}	1.4
15 years	4.2×10^{-11}	1.0
Adult	4.2×10^{-11}	1.0

dose to offspring. Data collected from studies in animals, and in humans where available, and were used as the basis for biokinetic models for the transfer of radionuclides to the fetus. These data also provide a basis for the relative concentrations of radionuclides averaged for the whole body of the fetus (C_F) and the mother (C_M). The ratios of $C_F:C_M$ are mainly taken

Tritium, Health Effects and Dosimetry. Table 9 ICRP Publication 88 [142]. Selected prenatal and infant dose coefficients

Tritiated compound type and mode of intake	Prenatal dose coefficients (Sv/Bq)		Dose coefficients for 3 month old (Sv/Bq)
	Acute maternal intakes ^a	Chronic maternal intakes ^b	
HTO inhalation	3.6×10^{-11}	3.1×10^{-11}	6.4×10^{-11}
HTO ingestion	3.6×10^{-11}	3.1×10^{-11}	6.4×10^{-11}
OBT ingestion	7.6×10^{-11}	6.3×10^{-11}	1.2×10^{-10}

^aValues are for acute intakes at the end of week 10 of pregnancy. Acute intakes occurring at other times yield lower dose coefficients

^bChronic intake begins at start of pregnancy and lasts for the duration of the pregnancy

from studies presenting data at short times postintake. The ICRP [142] report conservatively assumed that the recommended $C_F:C_M$ ratios applied to the time of intake and remained constant throughout pregnancy.

The ICRP [142] report presents dose coefficients for both chronic and acute maternal intakes by inhalation or ingestion. Acute intakes from 6 months before conception until the end of week 35 of pregnancy are considered. The report also provides chronic intakes lasting from 5 years or 1 year before conception until the time of conception, as well as chronic intakes lasting for the duration of the pregnancy. Effective doses to the offspring until birth, per unit intake by the mother, are presented, along with total committed effective doses to the offspring up to age 70, including doses resulting from radionuclides retained by the offspring at birth and doses received in utero.

HTO rapidly crosses the placenta after it is inhaled or ingested by the mother. Dose to the fetus resulting from such intakes depends on the water content of the fetus over the course of its development. As gestation progresses, water content decreases, while amounts of protein, fat, and minerals increase. The percentage of total body water decreases from about 93–95% at week 6 of pregnancy to 70–72% at birth. For comparison, the percentage of total body water in a nonpregnant woman is about 50%. While maternal total body water content is known to increase during pregnancy, studies disagree on the absolute values of total body water content. This disagreement is related to data on fetal body water content and the derivation of the amniotic fluid volume. However, calculations have shown that the maternal biological half-life of HTO varies from about 10 days at the start of pregnancy to

about 12 days at term. This variation in the biological half-life does not significantly affect the ICRP fetal dose coefficients for HTO [142].

Until week 8 of pregnancy, the dose to the embryo is taken to be equal to that of the uterus. This dose is proportional to the concentration of HTO in maternal body water. For dose to the fetus (after week 8), the HTO concentration in fetal body water is taken to be equal to that in the mother.

Using an average body water content of 80% for the fetus and 50% for the mother, the ICRP uses a $C_F:C_M$ ratio of 1.6 for HTO dose coefficients. This $C_F:C_M$ ratio of 1.6 is applied to both the HTO and OBT components for intakes of OBT. Fetal tritium is assumed to be uniformly distributed throughout all tissues. Following birth, the biokinetic model for the 3-month-old [127] is used to assess dose to the newborn. Table 9 presents selected prenatal dose coefficients with comparisons to dose coefficients for the 3-month-old.

These dose coefficients are greater than other chronic dose coefficients tabulated by the ICRP [142], namely for chronic intakes starting 5 years and 1 year before pregnancy and lasting until the start of pregnancy.

Nursing and Tritium Intake The ICRP [143] has recommended dose coefficients for newborns from intakes of radionuclides in maternal milk resulting from maternal intakes. As with prenatal doses, acute and chronic intakes by the mother are considered for intakes by both ingestion and inhalation. Nursing is assumed to continue for 6 months after birth, and ingestion dose coefficients for infants are applied [128]. Maternal intakes during pregnancy and during

Tritium, Health Effects and Dosimetry. Table 10 ICRP Publication 95 [143]. Selected nursing infant dose coefficients

Tritiated compound type and mode of intake	Dose coefficient (Sv/Bq)	
	Acute maternal intakes ^a	Chronic maternal intakes ^b
HTO inhalation	2.2×10^{-11}	2.0×10^{-11}
HTO ingestion	2.2×10^{-11}	2.0×10^{-11}
OBT ingestion	3.5×10^{-11}	3.0×10^{-11}

^aValues are based on acute maternal intake at 1 week after birth

^bAcute intakes occurring at other times yield lower dose coefficients

lactation are also considered. The ICRP's approach estimates the activity of radionuclides transferred to milk as a function of maternal intake for various intake scenarios (acute or chronic intake regimes, and for various maternal intake times relative to birth).

The HTO and OBT models described in previously were modified by the ICRP [143] to account for transfer to milk. The rate of OBT transfer to milk was taken to be that of carbon. The ICRP [143] presented dose coefficients based on the assumptions of 6 feeds per day and an average daily intake by the infant of 0.8 L. Table 10 shows selected dose coefficients for doses to nursing infants resulting from maternal intakes of tritium.

Review of Currently Available Information about the Biokinetics of Tritiated Compounds

Dosimetry of Tritiated Compounds Using the HTO Model The parameters in the ICRP's HTO biokinetic model described above include the water and carbon biological half-lives and the HTO/OBT partitioning coefficients. The bases for biological half-lives and the HTO/OBT partitioning coefficients are discussed below.

Several studies have examined the biological half-life of HTO in adults. Butler and Leroy [144] found this parameter to vary with water intake (decreasing with increasing water intake rate), ambient temperature (decreasing with increasing ambient temperature), and age (decreasing with increasing age in adults).

In this study, 310 cases of HTO intakes showed the biological half-life to vary from about 4 to 18 days, with a mean of 9.5 days. Other studies used fewer cases, but showed similar results for HTO [7]: from 6 days for eight cases [145] to 12 days for five cases [146]. The ICRP's HTO model adopted a biological half-life of 10 days for adults.

The biological half-life of 40 days that is used for the OBT compartment in body tissues is based on carbon turnover in the body. This was derived from the ratio of the Reference Man [5] carbon body content (16 kg) to the daily carbon intake (0.3 kg per day) and yields a biological half-life for carbon ($0.693 \times 16 \text{ kg} / 0.3 \text{ kg/day} = 37 \text{ days}$). The calculated value of 37 days was rounded to 40 days in the ICRP model.

Some studies have also reported a longer lived component, which represents retained tritium that has a biological half-life greater than water or that of carbon in the HTO model. Such a component contributes to the committed effective dose and may influence how bioassay data are interpreted. Example results from such studies are shown in Table 11. In general, OBT in compartment 3 has a long biological half-life relative to the half-life of 40 days used in the ICRP model for OBT. OBT in this long-lived compartment represents less than 1% of total OBT. Taylor [157] has taken this long-lived compartment into account in HTO modeling. This is discussed further later in this section.

The partitioning of HTO and OBT after intakes of tritium as HT was examined by Takeda and Kasida [158] as part of a study of the biokinetics of HTO in rats. These investigators found that "initially, the ratio of tissue-bound tritium to total tritium was about 3% in the kidney and 1–5% in other tissues." Based on this study, the ICRP model assumes that 3% of initial HTO intakes are partitioned to OBT. Other investigators have found values of less than 1%; for example, Snyder et al. [149] reported a value of 0.4%, based on a single human exposure case. A study of an unplanned acute HTO intake incident involving eight male workers suggested a long-term component (average $T_{1/2}$ of $74 \pm 18 \text{ days}$), which accounted for $0.5\% \pm 0.2\%$ of the initial HTO intake [156]. Balonov et al. [146] also reported a long-term component partitioning fraction of 0.5% for a human subject (compared to 3% in the ICRP HTO model).

Tritium, Health Effects and Dosimetry. Table 11 Selected tritium retention half-lives in humans following intakes of tritiated water

Study	Number of cases	Biological half-life (days)		
		Compartment 1 (body water)	Compartment 2 (organically bound)	Compartment 3 (organically bound)
Pinson and Langham [3]	9	11.3		
Foy et al. [147]	10	5–11 (mean: 7.5)		
Wylie et al. [148]	7	6.4–12.1 (mean: 8.5)		
Butler and Leroy [144]	310	4–18 (mean: 9.5)		
Osborne [137]	30	6.4–14.4 (mean: 10.5)		
Snyder et al. [149]	1	8.7	34	
Sanders and Reinig [150]	1	6.1 ^a	23	344
Minder [151]	1		10–30	139–230
Lambert et al. [152]	1	9.1 ^b	36	
Moghissi et al. [153] ^c	2		21 and 33	280 and 2,020 ^d
Moghissi et al. [154] ^e	3		21 and 26	280 and 550 350 ± 190 ^f
Bennet [155]		8.7	30	550
Balonov et al. [146]		12	39–76	
Rudran [145]	8	3.3–7.7 (mean: 6.0)	30.8–131 (mean: 81.7)	
Trivedi et al. [156]	8	6.2–12.8 ^g (mean: 8.4)	58–104 (mean: 74 ± 18)	

^aOral diuretic administered from days 3 to 35 postintake^bHT/HTO acute intake^cData for two tritium luminous dial painters collected 6–10 months after termination of employment^dLong-term behaviour of the data in one subject is attributed to seasonal effects, as described by Pinson [3]^eBased on additional data from subjects used in Moghissi et al. [153]^fData for a third subject, which was introduced in this study^gDuring the initial period when the exposed individuals increased fluid intakes, for 1 month postintake, the biological half-life varied from 5.0 to 8.1 days with a mean of 6.3 days

More recently, a three-compartment model was developed to assess HTO excretion data from the limited human data that have been published [157]. A three-component exponential function was derived

to represent the urinary excretion data from these studies. However, it was noted that results showed reasonably wide variations, both in observed biological half-lives and in proportions of total tritium entering

Tritium, Health Effects and Dosimetry. Table 12 Parameters for the HTO model recommended by Taylor [157]

Model component	Distribution (%)	Biological half-life (days)
HTO	99	10
OBT ^a	0.98	40
OBT ^b	0.02	350

^aShort-term OBT compartment

^bLong-term OBT compartment

the compartments (OBT), which turn over more slowly [157]. The model parameters are shown in Table 12. The resulting dose per unit intake of HTO by adults, based on this model, is 1.7×10^{-11} Sv/Bq. The current ICRP dose coefficient for HTO is 1.8×10^{-11} Sv/Bq [139]. This change would not be relevant for routine monitoring situations, where the dose is estimated by interpolating the tritium concentration in body water, between tritium-in-urine measurements taken frequently (for instance, every 2 weeks).

This latter method of dose estimation is independent of the time between intake and submission of a urine sample for tritium-in-urine measurement. Worker monitoring programs based on urine bioassay are typically used for routine monitoring when tritium doses are not significant, typically in cases that involve less than a few millisievert per year. Also, these programs are designed so that urine samples are submitted frequently enough to ensure accurate dose estimates. The method currently used to estimate doses from measurements of the concentration of tritium in urine has been described [129]. Workers' tritium-in-urine concentration is determined using liquid scintillation analysis. Linear interpolation is carried out between successive bioassay results, in order to calculate the workers' effective dose received between urine samples. Doses received during each annual monitoring period are summed to obtain the annual dose. The committed effective dose, resulting from intakes of tritium during the year, but received during the following year (as a result of tritium retention in the body) is also calculated by multiplying the year's last bioassay sample result by a committed dose factor [129, 131].

Taylor's model would be relevant for special bioassay monitoring following unplanned significant intakes

of HTO, where individual monitoring may extend several weeks postintake and when the individual has been removed from work that might result in further intakes of tritium.

Absorption of HTO Through Skin The absorption of tritium through skin from either the vapor phase or the liquid phase has been investigated by various authors, such as Pinson and Langham [3], DeLong et al. [4], and Osborne [137]. DeLong et al. [4] exposed mice, rats, and human adult volunteers (all males) to HTO vapor in air. Animals were sacrificed following exposure and urine samples were collected from human subjects for 48 h after the end of exposure. Water absorption rates, calculated from measurements of tritium in blood and total body water, suggest a lag in the distribution of the absorbed water. Absorption of tritium through the skin was the same when skin was covered with a cloth (cotton) as when it was not covered. The rate of water absorption was also found to be proportional to water vapor pressure. This suggests a single diffusion mechanism for percutaneous absorption. The data also suggests that the rate of percutaneous absorption of tritium from HTO in air is about the same as the rate of pulmonary absorption from the same atmosphere.

Pinson and Langham [3] exposed human subjects' forearm skin to HTO in the form of vapor as well as water. When exposed to HTO vapor, the average absorption rate was $0.018 \text{ mg/cm}^2/\text{min}$, compared with exposure to HTO in the liquid phase, where the absorption rate through skin was $0.04\text{--}0.065 \text{ mg/cm}^2/\text{min}$. The average absorption rate when exposed to HTO vapor was larger than could be accounted for by diffusion due to vapor pressure alone. A faster rate of uptake during short exposures was explained by a blotter effect (the ability of a material to draw a liquid into it). The study also found that the skin absorption rate increased with increasing skin temperature. As with DeLong [4], Pinson and Langham [3] found that "the quantity of HTO entering the body through the total skin, when exposed to an atmosphere containing a given activity per unit volume, would be about equal to that entering through the lungs."

Osborne [137] exposed volunteers (whole-body exposure) to HTO in air and measured tritium in urine. As with Pinson and Langham [3], Osborne

found a correlation between skin absorption rate and skin temperature. Overall, skin absorption rates were found to be similar to those reported by Pinson and Langham [3]. Osborne [137] also reported that the total skin intake rate assumed by the ICRP at that time [159] was twice the average intake rate observed. The assumption used in ICRP [159] was that the amount of tritium absorbed through the skin was equal to the amount inhaled. The assumption now is that the amount of tritium absorbed through the skin is half the amount inhaled.

The Inhalation of Elemental Tritium Gas The ICRP [134] recommended basing HT inhalation dose primarily on exposure of the lung, as opposed to exposure of the skin. Oxidation of HT to HTO in vivo was not considered at that time. The ICRP's Derived Air Concentration (DAC) [134] is calculated such that a typical worker breathing air, with a constant concentration of a specified radionuclide equal to the derived air concentration (DAC) for an entire working year, would receive a committed effective dose of 50 mSv. It should be noted the DAC is now based on a committed effective dose of 20 mSv. The DAC recommended at that time by the ICRP for HT was 2×10^{10} Bq/m³ compared with that for HTO, which was 8×10^5 Bq/m³.

A small fraction of inhaled elemental tritium gas is converted to HTO in the body. The absorption of HT through skin appears to be negligible and it does not convert to HTO on contact with skin [7]. Compared to HTO, HT is slightly soluble in body fluids and has a much lower uptake into biological systems. After inhalation, most HT is exhaled, but a small fraction is dissolved in body fluids and then oxidized to HTO by bacteria in the GI tract [160]. This is the only known biological site of HT oxidation [160].

The studies of Pinson and Laugham [3] and Peterman et al. [138] indicate that exposure to HT results in the excretion of HTO in urine, and that HTO formed from the oxidation of HT is retained in the body and excreted with the usual biological half-life of about 10 days. About 0.01% of HT gas inhaled by human volunteers was converted to HTO in the body [138].

Using HT and HTO biokinetic models proposed by Peterman et al. [138], in parallel with human volunteer HT inhalation data, it was concluded at the time that

the committed effective dose from inhaled HT was dominated by two roughly equal contributors: dose to lungs resulting from HT in inhaled air and dose from HTO formed by the oxidation of HT.

The ICRP has since recommended [139, 161] a revised dose coefficient for inhalation of HT gas. This stems from two sources: the work by Peterman et al. [138], described above, and the ICRP's revision of the human respiratory tract model – in particular, consideration of human respiratory tract morphology. The short range of tritium beta particles (6 μ m in tissue) is such that most of the energy is not deposited in target cells of the respiratory tract. The average cell depth of these cells' nuclei ranges from about 10 to 50 μ m in the extrathoracic, bronchial, and bronchiolar regions of the respiratory tract [139]. In the current ICRP [139] model of the respiratory tract, tritium beta particles are taken to deliver dose only in the alveolar-interstitial (AI) region after inhalation of HT. As discussed by Trivedi and Gentner [162], the nuclei of target cells within the AI region are at depths of less than 10 μ m and can be assumed to receive some dose from HT.

The lung dose rate calculated with the ICRP Publication 66 model [139] is about one third of the mean lung dose rate calculated with the ICRP Publication 30 [134] model. As a result, the effective dose per unit intake resulting from oxidized HT is several times higher than that due to the direct irradiation of the lungs by HT gas.

The revised ICRP HT dose coefficient, 1.8×10^{-15} Sv/Bq, is 10,000 times less than the coefficient for HTO [161]. This dose coefficient includes only the oxidation of HT into HTO, and not direct irradiation of the lungs by HT gas in air. The latter might increase the committed effective dose by about 20%. In cases where the dose is predominantly due to HT, exposure of the lungs should be taken into account.

It also follows from the above that the committed effective dose resulting from HT inhalation can be estimated using measurements of HTO concentration in urine [162].

An upper limit on the lung dose resulting from HT gas in the lungs can be calculated from HTO-in-urine measurements. However, this upper bound assumes that the entire intake occurred as HT. Some or all of the HTO in urine may have actually resulted from the

inhalation of HTO. Calculated in this manner, the contribution of lung dose to the committed effective dose is most likely overestimated, unless the HT to HTO ratio in air is known [162]. Consequently, when workers are exposed to both HTO and HT, the main contributor to dose is HTO and doses are not significant, it is adequate to calculate their dose from measurements of HTO in urine, without calculating an upper bound to lung dose from HT. When the dose is significant (e.g., more than a few millisievert for workers or close to the dose limit for members of the public), and due mainly to HT, exposure of the lungs should be taken into account by increasing the HT dose coefficient by 20%.

Intake Resulting from Skin Contact with Tritium-Gas-Contaminated Surfaces Studies have examined the behavior of tritium in the body as a result of skin contact with HT-contaminated metal surfaces (HT dissolves in most materials). These studies found that OBT from such skin contact exposures [163] is excreted in urine. In turn, the type of tritiated compound produced in the body as a result of the skin's contact with tritium will determine the resulting committed effective dose.

Elemental tritium gas and HTO are known to sorb to surfaces. There is negligible absorption of airborne elemental tritium gas through skin, but exposure of skin to tritium-gas-contaminated surfaces has been shown to result in the absorption of tritium – both as HTO and OBT [163]. When metal surfaces contaminated with tritium gas were applied to the inside forearm of four human volunteers, urinary excretion of tritium was found to reach a maximum rate about 24 h after exposure. In each subject, the excretion of total tritium at the peak time was about 20 times the amount of HTO excreted; that is, 5% of the tritium urinary excretion rate at the time of maximum excretion rate was due to HTO. The remaining tritium (that was not HTO) was determined to be organically bound. One to three weeks after the exposure, the rate of total tritium excreted was the same as that of HTO. Following intake via this route, up to 50% of the excreted OBT has been found to be excreted via urine following a biological half-life of about 1–2 days, and the rest of the OBT following a biological half-life of 0.1–0.2 days. HTO formed as a result of such skin

exposure to HT-contaminated surfaces has been found to be excreted in urine with a biological half-life of about 14 days.

About 0.5% of the tritium initially sorbed to the surface to which the volunteers were exposed was absorbed as HTO, and about 0.3% was absorbed as OBT. The results did not depend on the type of material, as long as the surface adsorbed tritium gas (similar results were obtained for various types of metals and for glass). In addition, the amount of tritium absorbed was found to be independent of the duration of exposure. The effective dose resulting from such intakes of tritium was estimated to range from 8.7×10^{-12} to 9.7×10^{-12} Sv/Bq absorbed [164].

Ingestion of Organically Bound Tritium While HTO distributes reasonably uniformly throughout body tissues, OBT formed in body tissues after intakes of OBT will be distributed among tissues according to their relative metabolic activities. OBT distribution will also depend on the chemical nature of the compound ingested. However, greater uptake by tissues with higher metabolic activity will generally be associated with shorter retention times, because of higher turnover. The nonuniform distribution of OBT is discussed in this section.

Experimental evidence indicates that, after an intake of OBT, about 9 times more tritium is bound to organic compounds in major tissues, compared with intakes of HTO [10]. Takeda and Kasida [158] also found that following HTO intakes in animals, 1–5% of tritium became incorporated in organic constituents of tissues. Therefore, 9–45% of tritium taken in as OBT would be expected to be incorporated in organic molecules in tissue. The ICRP's assumption that 50% of tritium remains as OBT after an intake of OBT stems from this reasoning.

The ICRP's HTO and OBT models [127] therefore predict that 17 times more tritium (50% vs 3%) will be incorporated as OBT following an acute OBT intake, compared to an HTO intake over a period of time (chronic intake). This is consistent with experimental results of Moghissi et al. [153], Rochalska and Szot [10], Sanders and Reinig [150], Snyder et al. [158], and Takeda and Kasida [149], which indicate ratios ranging from 3 to 20. It should be noted that the ICRP's OBT model predicts 43 times more OBT in

the body than does the Taylor [157] HTO model. However, Rodgers [15] found a ratio of about 12, suggesting that the current ICRP HTO model more accurately predicts the amount of OBT in the body from chronic intakes of HTO than does Taylor's HTO model.

The relative amounts of tritium incorporated into OBT following chronic intakes of HTO and OBT can be estimated from results of the study by Rodgers [15]. This study, carried out using mice, involved chronic intakes of HTO, OBT, or both for 56 days. At 56 days after the start of chronic intakes, the amounts of HTO and OBT in the mice would have been at steady state. The levels of OBT and HTO in the mice were then examined for about 1 month after the chronic exposure ended.

In the case of mice exposed to HTO only, the OBT concentration in tissues was about 22% of the HTO concentration. The ICRP HTO model [127] predicts this value to be about 12% at steady state, while the Taylor [157] model predicts this value to be about 5%.

In the case of mice exposed to OBT only, the OBT concentration in tissues was about a factor of 6 greater than the HTO concentration. This compared to the factor of 4 predicted by the ICRP OBT model [127].

This study also found that the amount of OBT in tissues resulting from OBT intake was about 12 times greater than the amount of OBT in tissues resulting from HTO intakes, which is in close agreement with the results from Rochalska and Szot [10]. Although the ICRP models appear to underestimate the amounts of OBT retained in the body in some cases of chronic intakes, they predict that the dose attributable to OBT, from intakes of OBT, is 10 times greater than the dose from OBT from intakes of HTO. This is in agreement with Rodgers' findings. Overall, it can be concluded that these studies are generally consistent and that the ICRP model does not underestimate dose.

Rogers [15] observed that after chronic exposure had ended, OBT was lost from the mice at a rate best described by a two-compartment model, with about 40% of OBT being cleared rapidly (with half-lives of 2.0–3.2 days in mice), and the remaining OBT being cleared much more slowly (with half-lives of 24–30 days in mice). The faster-clearing compartment's rate constant was not significantly different from those associated with HTO clearance from mice. Rogers

attributed this faster-clearing compartment to loosely bound or exchangeable OBT. The slower-clearing compartment was due to nonexchangeable OBT.

Rodgers [15] also found that when exposed to HTO only, the HTO in the mice contributed more than 95% of the total tritium dose during the chronic exposure. In the case of mice exposed to OBT only, the study found that about 50% of the total dose was attributable to OBT in tissue (in adult humans, the HTO:OBT dose ratio follows the ratio of the half-lives; 10 days:40 days). However, after the end of chronic intake, due to the longer biological half-life of OBT, the total dose resulting from OBT (in the case of mice exposed to OBT only) was about one order of magnitude greater than that of mice exposed to HTO only. This dominance of the dose attributable to OBT began after the end of chronic intake, as HTO cleared faster than OBT. The latter is consistent with predictions of the ICRP's OBT model; that is, a dose from OBT that is one order of magnitude greater after the end of chronic intake.

In order to relate the study by Rodgers [15] to human dietary intakes of tritium, a study by Trivedi et al. [135] is worthy of consideration. OBT:HTO ratios greater than 1 have been observed in some foods grown near environmental sources of tritium. The range of the OBT:HTO ratio in dietary intakes varies according to many factors, such as diet and atmospheric conditions. A study of doses received by a population residing near a nuclear facility showed that OBT contributes about 16% of the total tritium dose from all sources of tritium exposure [165]. In this case, the average OBT:HTO ratio in food items consumed by these individuals was 0.04. This value was mainly influenced by the proportion of food that was not locally grown, but that was imported into the study region. From this information, Trivedi and Gentner [162] concluded that "unless the population living near tritium facilities consumes locally grown food items for a large proportion of the year (an uncommon situation around Canadian CANDU stations), dietary intake of OBT has only minor contributions to dose."

There are some uncertainties associated with OBT metabolism as described by the ICRP's OBT model, particularly the retention of OBT in children and the effect of organ growth on OBT retention. Also, the relative depositions of OBT in various organs and tissues are not fully known for all age groups. This implies

there is no age-dependent biokinetic model for dietary intakes of OBT that accounts for age-related physiological and anatomical variations. HTO turnover in the human body is well established, but measurement data for tritium turnover in organic compounds are limited, particularly for long-term biokinetic components.

Animal studies have shown a nonuniform distribution of OBT among soft tissues of the body. Also, within a single organ, tritium is distributed nonuniformly among the various organic molecules [127]. However, other than for the special case of DNA precursors it appears that any nonuniformity of distribution within cells will be small, and in any case will be reflected in RBE values.

Several studies have investigated the distribution of OBT within cells and the bodies of test animals. These have helped produce dosimetry models of energy deposition within cells. The section on “[Electron Energy and Ionization Path Lengths: Why Tritium Radiation Is More Effective Than Gamma Radiation](#)” provides more details on dosimetry aspects, but the work by Chen [121] warrants discussion here as it estimates an RBE for tritium. Chen [121] ran a Monte Carlo simulation of beta radiation energy tracts from tritium bound to radiosensitive sites with varying spherical diameters (10 nm to 2 μm) and compared this to the energy distribution of HTO that was homogeneously distributed throughout the body. As expected, the “dose mean linear energies” of OBT were higher than that of the HTO by a factor of 1.7, which was close to the value found by Straume and Carsten [25]. However, as the AGIR [27] points out, “the extent of any increase will depend on the extent that OBT does preferentially localize within critical targets”; that is, if the OBT were homogeneously distributed, then the added effect would be much lower.

About 70% of the hydrogen in the human body is in the exchangeable form, with about 95% of this hydrogen forming water molecules. Nonexchangeable hydrogen forms covalent bonds with carbon, and is found in components such as protein, lipids, and carbohydrates. Several factors determine the distribution of OBT in the body resulting from tritium in the diet. These include the ratio of OBT to HTO in the dietary items; daily intakes of carbohydrates, fats, and proteins; oxidation rates of the dietary constituents; the gut absorption factors for organic precursors

(namely monosaccharides, fatty acids, and amino acids); the retention of organic precursors; and HTO and OBT excretion rates [166].

Physiological models of OBT biokinetics have been proposed by Richardson [167], Richardson and Dunford [168], and Melintescu et al. [22]. They provide sex-differentiated biokinetics and dose coefficients for males and females. Richardson et al. [169] present sex differences in dose coefficients based on the ICRP model, which stem from a daily carbon intake rate that differs between females (228 g carbon) and males (303 g carbon). The resulting biological half-life of carbon is 51 days for female adults and 40 days for male adults. Richardson and Dunford [168] consider the differing water and organic contents of tissues and organs, as well as differing biokinetics for their organic components, as presented in the non-ICRP “HCNO Model” [168]. Table 13 shows dose per unit intake values for adults from these studies and the values as published by the ICRP for its HTO and OBT models. It can be seen that dose coefficients from the various

Tritium, Health Effects and Dosimetry. Table 13 Adult dose coefficients from physiologically based OBT biokinetic models

Biokinetic model	Dose coefficient (Sv/Bq)
Ingestion of HTO	
ICRP 72 [141]	1.8×10^{-11}
Richardson et al. [161]	1.8×10^{-11} (males)
	2.2×10^{-11} (females)
Melintescu et al. [22]	1.6×10^{-11} (males)
	2.3×10^{-11} (females)
Ingestion of OBT	
ICRP 72 [141]	4.2×10^{-11}
Richardson et al. [166]	4.2×10^{-11} (males)
	6.1×10^{-11} (females)
Richardson [167]	7.6×10^{-11} (adults, PNM model)
Richardson and Dunford [168]	7.4×10^{-11a}
Melintescu et al. [22]	4.7×10^{-11} (males)
	7.0×10^{-11} (females)

^aFor “Reference Man” diet [5]

studies in Table 13 are within a factor of 2 of the ICRP [141] values, with the highest one from Richardson and Dunford [168]. It should be noted that Richardson and Dunford [168] also provide dose coefficients for tritiated nutrients, with the highest value for tritiated protein at 8.4×10^{-11} Sv/Bq, which is twice the ICRP [141] value for OBT.

Richardson et al. [166] conducted a literature review of studies on OBT exposure and found that there exists a significant amount of information on intakes of tritiated food from animal experiments, but very little data from human intakes. The review also found there is sparse information on the exchangeable and nonexchangeable forms of OBT in diet. This, in turn, may have introduced uncertainties in current dosimetric models for tritium in the diet. Based on this literature review, Richardson [167] proposed a biokinetic model based on the overall catabolic reactions of foodstuffs, the Principal Nutrient Metabolic (PNM) model. This model accounts for the nonuniform deposition of OBT in tissues and organs, but, as with the ICRP model, does not include compartments representing the long-term retention of structural proteins and fat as adipose tissue.

It can be seen from Table 13 that taking gender differences into account increases the dose coefficient by up to about 25% for HTO and up to 50% for OBT. Accounting for the nonuniform distribution of OBT increases the dose by up to 80%.

Biokinetics of Tritiated DNA Precursors Tritium may be incorporated into DNA or RNA for the purpose of experiments studying cell kinetics. This section examines the biokinetics of such compounds. The nucleic acid precursors thymidine and deoxycytidine are most commonly labeled with tritium for this purpose. Uridine and adenine are most commonly labeled with tritium for work involving RNA [27].

After tritiated thymidine ($^3\text{HTdR}$) is administered to humans or animals, it is incorporated into DNA during the synthesis stage of the cell cycle. Following ingestion of $^3\text{HTdR}$, about 2% of the tritium is incorporated into DNA [49], and the remainder appears as HTO. Tritiated thymidine is available for only a short time after intake and primarily for uptake by rapidly cycling cells such as bone marrow or gut. Chronic exposure to $^3\text{HTdR}$ will also label slowly cycling cells.

Hence, the rate of turnover of these cells will determine the retention rate of this tritium.

The NCRP [8] examined the dose resulting from ingestion of $^3\text{HTdR}$ based on considerations that are mainly theoretical. In the case of acute intakes by ingestion, it was concluded that tritium ingested as thymidine is 8.6 times more hazardous than that ingested as HTO. It also suggests this may need revision as more precise data on stem cells and $^3\text{HTdR}$ distribution and incorporation rates become available. Conclusions regarding chronic intakes are similar. In the cases of other tritium-labeled DNA and RNA precursors, the recommendation is to apply exposure limits for $^3\text{HTdR}$ until more data on the radiation effects of these other compounds become available. $^3\text{HTdR}$ is unstable above -4°C and therefore may degrade once released into the environment.

The NCRP [8] states

- Because DNA is metabolically stable, in contrast to RNA, label bound to RNA precursor will be distributed, as a function of time, with increasing preference to DNA. Hence, for cells with a long lifespan, it is finally the rate of cellular proliferation that defines the biological half-life of the radionuclide incorporated in the cell nucleus, irrespective of whether it was initially bound to RNA or DNA. On this basis, the dose to the stem cell nucleus from ^3H -thymidine is about 50 times higher than from the same amount administered as ^3HOH

Furthermore, the same report states

- it appears justified to estimate the long-term hazard from labeled RNA precursors . . . to be not larger than that from the same amount of equally labeled thymidine. Until more experimental evidence is available, late hazards from labeled RNA precursors should be considered similar to those from equally labeled thymidine

Komatsu et al. [170] examined the dose resulting from tritium bound to DNA, as a result of ingesting HTO and OBT. These investigators fed HTO and food to mice for 22 days. Equilibrium was reached by the 16th day of chronic intake. In the group of mice fed tritiated food, the highest ratio of OBT to HTO was found to be 2.2 in the liver. Liver DNA was found to have a lower specific activity than that of liver OBT. This was attributed to tritiated DNA containing little, if any, exchangeable OBT,

which has a high specific activity. Comparing the specific activities of DNA-bound tritium in both groups of mice (those fed HTO and those fed tritiated food) the specific activity of DNA-bound tritium in the tritiated food group was found to be 4.6 times that of the HTO group. The study also suggests that the ratio of the specific activity of DNA-bound tritium to HTO may be higher in embryos than in adults.

Tritium bound to DNA was found to contribute little (less than 1%) to the total dose as a result of chronic exposure to either HTO or food. The total doses for both exposure groups were found to be similar for chronic intakes. In the case of acute intakes of HTO, DNA-bound tritium was found to contribute about 20% to the total dose from intakes of HTO and about 40% to the total dose from intakes of tritiated food. The total dose resulting from the ingestion of tritiated food was found to be about double that from the ingestion of HTO.

More recently, Melintescu et al. [22] reported that tritium can be strongly bound in the hydration shell, mostly in DNA. This tritium, referred to as buried tritium, represents about 5% of the OBT concentration in plants and 20% of that in fish. Buried tritium contributes to the total dose resulting from tritium intake. Melintescu et al. [22] report that effects due to buried tritium are included in the assessment of the RBE and radiation weighting factor, so dose from buried tritium does not need to be added to the total tritium dose calculation.

Arguments have been presented suggesting that tritiated DNA and histone precursors present a significantly greater risk than that resulting from HTO [171, 172]. The author indicates such tritiated compounds may be 1,000–5,000 times more effective than HTO because of their heterogeneous distribution in the body. The author attributes this mainly to the short range of tritium's beta particles, which causes their energy to be deposited in the cell nucleus. The authors also note that it remains to be proven that no significant risk exists from tritium bound to DNA as a result of releases of HTO to the environment (from HTO taken up by plants and animals, converted into OBT by these plants or animals, ingested by humans and incorporated into human cell nuclei).

Bridges [173, 174] responded to both articles by Müller [171, 172] by stating that the approximate

1,000-fold difference in effects is based on units of activity (such as becquerels) in the medium, rather than in terms of the dose to the cell or the nucleus. Also, the work referred to by Müller was carried out on *in vitro* systems, while *in vivo* results are quite different. Almost all ingested thymidine is broken down in the gut, and there is evidence that the effectiveness of $^3\text{HTdR}$ in adult mammals is only slightly greater than that expected from HTO. Furthermore, free $^3\text{HTdR}$ is unstable in the environment and therefore unlikely to be a source of exposure to tritiated thymidine. Available information indicates that the amount of tritium incorporated into DNA in *in vivo* systems is too small to cause the RBE to increase.

Summary and Conclusions

While current ICRP models for intake of HTO, HT, and OBT are reasonably consistent with experimental results, further work related to HT and OBT biokinetics is suggested in this section.

HTO behavior is well understood, and the current ICRP model appears to accurately predict HTO body burden and urinary excretion. The ICRP is considering adopting a revised biokinetic model for HTO [157]. Although the Taylor [157] model for HTO has a dose coefficient similar to that of the current ICRP model, it treats OBT differently by considering the long-term retention of tritium. It has been proposed that the ICRP adopt the Taylor Mode. However, given the variation in the limited data on which the Taylor model was based, it would be beneficial to assess it against further human data. Moreover, it would be of interest to compare the Taylor model's predicted contribution of OBT to dose with the work of Trivedi et al. [135, 136], which studied OBT contribution to dose in actual human HTO exposure cases. Finally, the Taylor model applies only to adults; an expansion to various age groups is needed to satisfy the requirements of public dose assessments.

When elemental tritium gas is inhaled, the dose to the cells of the respiratory tract's AI region should be considered when the exposure is predominantly to HT.

The ICRP's OBT model appears to be generally consistent with experimental results. However, it does not account for the differentiated deposition of OBT between various organs and tissues. The use of

physiologically based OBT models, such as the one proposed by Richardson and Dunford [168], should be encouraged when significant doses are expected (such as in nonroutine or accidental exposures). The validation and incorporation of these models in computer codes would facilitate their use in OBT dose assessments. Furthermore, it would be valuable to expand this type of model to include age dependency for all age groups, including the nursing infant.

The NCRP [8] presented a review and interpretation of data on the distribution of radionuclides incorporated in genetic material. A review of recent work on this topic would allow this information to be updated. Finally, Komatsu et al. [170] concluded that the total dose resulting from the ingestion of tritiated food is about twice that from the ingestion of HTO. This suggests that current OBT dose coefficients account for the binding of tritium to DNA.

Future Directions

Tritium is a significant by-product from nuclear reactors that use heavy water as a neutron moderator. Tritium releases from such plants form the largest contribution to radiation dose from nonnatural sources to biota in the environment and to members of the public.

While the biological effects of tritium have likely been studied more than that of any other radionuclide, some uncertainty still remains with respect to effects at lower doses. This has consequences both in radiological protection of humans and the environment. As illustrated in this report, the outcome of most concern at low doses is that of stochastic effects such as cancer and possibly hereditary effects.

The study of stochastic effects is challenging due to limitations in conventional animal models (e.g., short life spans). These studies are especially important because there is little information known about the RBE of tritium radiation when the endpoints are statistically based, such as cancer. Newer breeds of laboratory animals, such as those more susceptible to some cancers, may aid in meeting those challenges. Also, advances in molecular biology may provide answers at the cellular level.

Furthermore, questions still remain on the effectiveness of various OBT compounds as well as how they behave within the body. For example, a tritiated fat

molecule will have a much longer residency within the body, but its radiological consequence would be much less than a tritiated DNA precursor with a much shorter biological residency. Therefore, more work is needed in first identifying how the various OBT compounds are partitioned within the body and then to understand the radiological consequences.

Similarly, epidemiological studies of humans occupationally exposed to tritium may yield some information on health effects. To date, such studies are complicated by small numbers in exposed populations or confounding factors such as concurrent exposures to other radiations or chemical carcinogens. An international effort involving the grouping of populations would be required to attain suitable statistical power.

Abbreviations

AECL	Atomic Energy Canada Limited
AGIR	Advisory Group on Ionizing Radiation
AI	Alveolar-interstitial
BEIR	Biological Effects of Ionizing radiation
BW	Body weight
C_F	Concentration fetus
C_M	Concentration mother
CANDU	Canadian deuterium uranium
CERRIE	Committee Examining Radiation Risk from Internal Emitters
CI	Confidence interval
CNSC	Canadian Nuclear Safety Commission
CTH₃	Tritiated methane
DAC	Derived air concentration
DDREF	Dose and dose rate effectiveness factor
DNA	Deoxyribonucleic acid
GI	Gastro-intestinal
H₂O	Chemical symbol for water, also seen as HOH
HCNO	Hydrogen carbon nitrogen oxygen
HSE	Health and Safety Executive
HT	Tritiated gas
³HTdR	Tritiated thymidine
HTO	Tritiated water
ICRP	International Commission on Radiological Protection
ICRU	International Commission on Radiological Units and Measurements
LD	Lethal dose

LD_{50/30}	Lethal dose that kills 50% of the population in 30 days
LDEF	Low dose effectiveness factor
LET	Linear energy transfer
N	Number
NCRP	National Council on Radiation Protection
NGS	Nuclear generating station
NPP	Nuclear power plant
OBT	Organically bound tritium
OH	Hydroxyl
OPG	Ontario Power Generation
OR	Odds ratio
PNM	Principal nutrient metabolic
PSL	Priority substance list
PWR	Pressurized water reactor
Q	Quality factor
RBE	Relative biological effectiveness
RBE_{MAX/M}	Maximum relative biological effectiveness
REF	Radiation effectiveness factor
RR	Relative risk
SMR	Standardized mortality ratio
SSB	Single strand break
SSK	German Commission on Radiological Protection
Type F,M,S	Type fast, medium, slow, absorption rates
UK	United Kingdom
UKAEA	United Kingdom Atomic Energy Authority
UNSCEAR	United Nations Scientific Committee on the Effects of Atomic Radiation
w_R	Radiation weighting factor
w_T	Tissue weighting factor
χ²	Chi squared

Units

Bq: becquerel
 cm: centimeter
 d: day
 g: gram
 GBq: gigabecquerel
 Gy: gray
 h: hour
 kBq: kilobecquerel
 keV: kiloelectron volts

kg: kilogram
 km²: kilometer squared
 kV: kilovolt
 kVcp: kilovolt constant potential
 kVp: kilovolt peak
 L: liter
 MBq: megabecquerel
 m: meter
 m²: meter squared
 m³: meter cubed
 mGy: milligray
 mL: milliliter
 mm: millimeter
 mSv: millisievert
 R: röntgen
 Sv: sievert
 μSv: microsievert
 μm: micrometer
 y: year

Bibliography

1. Carsten AL (1979) Tritium in the environment: isotopic effects and transmutation. In: Lett JT, Adler H (eds) *Advances in radiation biology*. Academic, New York
2. Virsik RP, Schäfer CH, Harder D (1980) Chromosome aberrations induced in human lymphocytes by ultrasoft Al_K and C_K X-rays. *Int J Radiat Biol* 38(5):545–557
3. Pinson EA, Langham WH (1957) Physiology and toxicology of tritium in Man. *J Appl Physiol* 10:108–126
4. DeLong CW, Thompson RC, Kornberg HA (1954) Percutaneous absorption of tritium oxide. *Am J Roentgenol Radium Ther Nucl Med* 71:1038–1045
5. International Commission on Radiological Protection (ICRP) (1975) Report of the task group on reference man. ICRP Publication 23. Pergamon, New York
6. Belloni P, Clemente GF, Di Piero S, Ingrao G (1983) Tritium levels in blood and urine samples of the members of the Italian general population and some exposed subjects. *Radiat Prot Dosim* 4:109–113
7. Hill RL, Johnson JR (1993) Metabolism and dosimetry of tritium. *Health Phys* 65(6):628–647
8. National Council on Radiation Protection and Measurements (NCRP) (1979) Tritium and other radionuclide labelled organic compounds incorporated into genetic material. Report No. 63 U.S. NCRP, Washington, DC
9. Amano H, Atarashi M, Noguchi H, Yokoyama S (1995) Formation of organically bound tritium in plants during the 1994 chronic HT release experiment at Chalk River. *Fusion Technol* 28:814–820
10. Rochalska M, Szot Z (1977) The incorporation of organically-bound tritium of food into some organs of the rat. *Int J Radiat Biol* 31(4):391–395

11. Saito M, Ishida MR (1989) Dose-modification factor for accumulated dose to cell nucleus due to protein-bound H-3. *Health Phys* 56:869–874
12. Diabaté S, Strack S (1993) Organically bound tritium. *Health Phys* 65(6):698–712
13. United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) (2008) United Nations. Effects of ionizing radiation. Volume II, Annex C – Non-targeted and delayed effects of exposure to ionizing radiation. 2006 Report to the General Assembly, with scientific annexes. United Nations, New York
14. Persaud R, Zhou H, Baker SE, Hei TK, Hall EJ (2005) Assessment of low linear energy transfer radiation-induced bystander mutagenesis in a three-dimensional culture model. *Cancer Res* 65:9876–9882
15. Rodgers DW (1992) Tritium dynamics in mice exposed to tritiated water and diet. *Health Phys* 63:331–337
16. Richmond CR, Langham WH, Trujillo TT (1962) Comparative metabolism of tritiated water by mammals. *J Cell Comp Physiol* 59:45–53
17. Brues AM, Stroud AN, Rietz L (1952) Toxicity of tritium oxide to mice. *Proc Soc Exp Biol Med* 79(1):174–176
18. Furchner JE (1957) Relative biological effectiveness of tritium beta-particles and cobalt 60 gamma-rays measured by lethality in CF1 mice. *Radiat Res* 6:483–490
19. Yokoro K, Yamamoto O, Seyama T, Kinomura A (1986) Acute and chronic effects of tritiated water in mice with special reference to its carcinogenicity: an interim report. *Radiat Prot Dosim* 16:165–168
20. Yamamoto O, Yokoro T, Seyama A, Kinomura T (1990) HTO oral administration in mice. I: threshold dose rate for haematopoietic death. *Int J Radiat Biol* 57(3):543–549
21. Soloviev V, Ilyin LA, Baranov AE et al (2001) In: Gusev IA, Guskova AK, Mettler FA (eds) Medical management of radiation accidents, 2nd edn. CRC Press, Boca Raton, pp 46 and 158
22. Melintescu A, Galeriu D, Takeda H (2007) Reassessment of tritium dose coefficients for the general public. *Radiat Prot Dosim* 127(1–4):153–157
23. National Research Council (NRC) (1980) The effects on populations of exposure to low levels of ionizing radiation. The committee on the biological effects of ionizing radiations (BEIR III). The National Academies Press, Washington, DC
24. National Research Council (NRC) (2006) Health Risks from exposure to low levels of ionizing radiation: BEIR VII Phase 2. Board on radiation effects research. The committee on the biological effects of ionizing radiations (BEIR). The National Academies Press, Washington, DC
25. Straume T, Carsten AL (1993) Tritium radiobiology and relative biological effectiveness. *Health Phys* 65(6):657–672
26. Commerford SL, Carsten AL, Cronkite EP (1982) The turnover of tritium in cell nuclei, chromatin, DNA, and histone. *Radiat Res* 92:521–529
27. Health Protection Agency (HPA) (2007) Review of risks from tritium with particular attention to tritiated water and organic compounds containing tritium. Report of AGIR subgroup on tritium risks, UK advisory group on ionizing radiation. http://www.hpa.org.uk/radiation/advisory_groups/agir/index.htm
28. Brown RM (1995) The measurement of tritium in Canadian food items. Atomic Energy Control Board (now the Canadian Nuclear Safety Commission) AECB INFO 0499. Atomic Energy Control Board, Ottawa
29. Dekaban AS (1968) Abnormalities in children exposed to x-radiation during various stages of gestation: tentative timetable of radiation injury to the human fetus. *Int J Nucl Med* 9(9):471–477
30. Mulvihill JJ, Harvey EB, Boice JD Jr, Chakravarti A, Miller RW (1991) Normal findings 52 years after in utero radiation exposure. *Lancet* 338(8776):1202–1203
31. Zamenhof S, van Martheus E (1979) The effects of chronic ingestion of tritiated water on the prenatal brain development. *Radiat Res* 77:117–127
32. Laskey JW, Pamsh JL, Cahill DF (1973) Some effects of lifetime parental exposure to low levels of tritium on the F2 generation. *Radiat Res* 56(17):1–179
33. Bursian SJ, Cahill DF, Laskey JW, Parker LN (1975) Some aspects of brain neurochemistry after intrauterine exposure to tritium. *Int J Radiat Biol Relat Stud Phys Chem Med* 27(5):455–461
34. Jones DC, Krebs JS, Sasmore DP, Mitoma C (1980) Evaluation of neonatal squirrel monkeys receiving tritiated water throughout gestation. *Radiat Res* 83(3):592–606
35. Dobson RL, Cooper MF (1974) Tritium toxicity: effects of low-level $^3\text{H}_2\text{O}$ exposure on developing female germ cells in the mouse. *Radiat Res* 58:91–100
36. Dobson RL, Kwan TC (1976) The RBE of tritium radiation measured in mouse oocytes: increase at low exposure levels. *Radiat Res* 66:615–625
37. Lushbaugh CC, Ricks RC (1972) Some cytokinetic and histopathologic considerations of irradiated male and female gonadal tissues. In: Vaeth JM (ed) *Frontiers of radiation therapy and oncology*, vol 6. Karger, Basel, pp 228–248
38. Lushbaugh CC, Casarett GW (1976) The effects of gonadal irradiation in clinical radiation therapy: a review. *Cancer* 37(2 Suppl):1111–1125
39. Baker TG, Neal P (1977) Action of ionizing radiations on the mammalian ovary. In: Zuckerman S, Weir BJ (eds) *The ovary*, vol 3. Academic, New York
40. United Nations Scientific Committee on the effects of ionising radiation (UNSCEAR) (1982) Ionising radiation: sources and biological effects, Vienna
41. Baker TG (1971) Comparative aspects of the effects of radiation during oogenesis. *Mutat Res* 11:9–22
42. International Commission on Radiological Protection (ICRP) (1984) A compilation of the major concepts and quantities in use by the ICRP. ICRP Publication 42. Ann ICRP 14(4). Pergamon, New York
43. National Research Council (NRC) (1990) Health risks from exposure to low levels of ionizing radiation. The committee on the biological effects of ionizing radiations BEIR V. National Academies Press, Washington, DC

44. Satow Y, Hori H, Lee J-Y, Ohtaki M, Sawada S, Nakamura N, Okada S (1989) Effect of tritiated water on female germ cells: mouse oocytes killing and RBE. *Int J Radiat Biol* 56(3):293–299
45. Satow Y, Hori H, Lee J-Y (1989) Teratogenic effect of fission neutron and tritium water on rat embryo. *J UOEH* 11(Suppl):416–431
46. Oakberg EF (1955) Sensitivity and time of degeneration of spermatogenic cells irradiated in various stages of maturation in the mouse. *Radiat Res* 2:369–391
47. Oakberg EF (1959) Initial depletion and subsequent recovery of spermatogonia of the mouse after 20 R of gamma rays and 100, 300, and 600 R of x-rays. *Radiat Res* 11:700–719
48. Oakberg EF, Clark E (1964) Species comparison of radiation response of the gonads. In: Carlson WD, Gaffner FX (eds) *Effects of ionizing radiation on the reproductive system*. Pergamon, Oxford, pp 11–24
49. Lambert BE (1969) Cytological damage produced in the mouse testes by tritiated thymidine, tritiated water and X-rays. *Health Phys* 17:547–557
50. National Research Council (NRC) (1988) Health risks of radon and other internally deposited alpha-emitters. The committee on the biological effects of ionizing radiations (BEIR IV). National Academy Press, Washington, DC
51. United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) (2000) United Nations. Sources and effects of ionizing radiation, vol I: sources; vol II: effects. Report to the General Assembly, with scientific annexes. United Nations sales publication E.00.IX.3 and E.00.IX.4. United Nations, New York
52. Tanaka IB, Tanaka S, Ichinohe K, Matsushita S, Matsumoto T, Otsu H, Oghiso Y, Sato F (2007) Cause of death and neoplasia in mice continuously exposed to very low dose rates of gamma rays. *Radiat Res* 167(4):417–437
53. Yamamoto O, Seyama T, Jo T, Terato H, Saito T, Kinomura A (1995) Oral administration of tritiated water (HTO) in mouse. II. tumour development: I/II/III. *Int J Radiat Biol* 68(1):47–54
54. Yamamoto O, Seyama T, Itoh H, Fujimoto N (1998) Oral administration of tritiated water (HTO) in mouse. III: low dose-rate irradiation and threshold dose-rate for radiation risk. *Int J Radiat Biol* 73(5):535–541
55. Gragtmans NJ, Myers DK, Johnson JR, Jones AR, Johnson LD (1984) Occurrence of mammary tumors in rats after exposure to tritium beta rays and 200 kVp x-rays. *Radiat Res* 99(3): 636–650
56. Johnson JR, Myers DK, Jackson JS, Dunford DW, Gragtmans NJ, Wyatt HM, Jones AR, Percy DH (1995) Relative biological effectiveness of tritium for induction of myeloid leukemia in CBA/H mice. *Radiat Res* 144(1):82–89
57. Seyama T, Yamamoto O, Kinomura A, Yokoro K (1991) Carcinogenic effects of tritiated water (HTO) in mice: in comparison to those of neutrons and gamma-rays. *J Radiat Res* 32(Suppl 2):132–142
58. Balonov MI, Muskinova KN, Mushkacheva GS (1993) Tritium radiobiological effects in mammals: review of experiments in the last decade in Russia. *Health Phys* 65(6):713–726
59. Mushkacheva GS, Rusinova GG, Muskinova KN, Lemberg VK, Shorokhova VB, Revina VS, Uryadnitskaya TI (1992) In: Buldakov L (ed) *Genetic structures and tritium*. Energoatomizdat, Moscow (in Russian)
60. Straume T (1993) Tritium risk assessment. *Health Phys* 65(6):673–682
61. International Commission on Radiological Protection (ICRP) (1991) 1990 recommendations of the international commission on radiological protection. ICRP Publication 60, Ann ICRP 21(1-3):1–201. Pergamon, Oxford
62. United Nations Scientific Committee on the effects of ionizing radiation (UNSCEAR) (1988) Sources, effects and risks of ionizing radiation, Vienna
63. International Commission on Radiological Protection (ICRP) (2007) Recommendations of the ICRP. ICRP Publication 103. Ann ICRP 37(2-4). Pergamon, Oxford
64. United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) (2001) Hereditary effects of radiation. Report to the General Assembly, with Scientific Annex. United Nations, New York
65. Furuno-Fukushi I, Ueno AM, Matsudaira H (1987) Cell killing and mutation to 6-thioguanine resistance after exposure to tritiated amino acids and tritiated thymidine in cultured mammalian cells (L5178Y). *Radiat Res* 110:428–438
66. Ueno AM, Furuno-Fukushi I, Matsudaira H (1982) Induction cell killing, micronuclei, and mutation to 6-thioguanine resistance after exposure to low-dose-rate γ rays and tritiated water in cultured mammalian cells (L5178Y). *Radiat Res* 91:447–456
67. Wang B, Watanabe K, Yamada T, Takeshi S, Shima A (1996) Effects of beta radiation from organically bound tritium on cultured mouse embryonic mid brain cells. *Health Phys* 71(6):915–921
68. Trivedi A, Morrison DP, Gentner NE (1997) Relative biological effectiveness for organically bound tritium. *Health Phys* 73(2):397–398
69. Steel GG, Lamerton LF (1965) The turnover of tritium from thymidine in tissues of the rat. *Exp Cell Res* 37:117–123
70. Lambert BE, Clifton RJ (1968) Radiation doses resulting from the ingestion of tritiated thymidine by the rat. *Health Phys* 15:3–9
71. Preston DL, Shimizu Y, Pierce DA, Suyama A, Mabuchi K (2003) Studies of mortality of atomic bomb survivors. Report 13: solid cancer and noncancer disease mortality: 1950–1997. *Radiat Res* 160:381–407
72. Preston DL, Pierce DA, Shimizu Y, Cullings HM, Fujita S, Funamoto S, Kodama K (2004) Effect of recent changes in atomic bomb survivor dosimetry on cancer mortality risk estimates. *Radiat Res* 162:377–389
73. Preston DL, Ron E, Tokuoka S, Funamoto S, Nishi N, Soda M, Mabuchi K, Kodama K (2007) Solid cancer incidence in atomic bomb survivors: 1958–1998. *Radiat Res* 168(1):1–64
74. Boice JD Jr, Day NE, Andersen A, Brinton LA, Brown R, Choi NW, Clarke EA, Coleman MP et al (1985) Second cancers following radiation treatment for cervical cancer. An international

- collaboration among cancer registries. *J Natl Cancer Inst* 74: 955–975
75. Boice JD Jr, Engholm G, Kleinerman RA, Blettner M, Stovall M, Lisco H, Moloney WC et al (1988) Radiation dose and second cancer risk in patients treated for cancer of the cervix. *Radiat Res* 116:3–55
 76. Weiss HA, Darby SC, Doll R (1994) Cancer mortality following X-ray treatment for ankylosing spondylitis. *Int J Cancer* 59:327–338
 77. Weiss HA, Darby SC, Fearn T, Doll R (1995) Leukemia mortality after X-ray treatment for ankylosing spondylitis. *Radiat Res* 142:1–11
 78. Howe GR (1995) Lung cancer mortality between 1950 and 1987 after exposure to fractionated moderate-dose-rate ionizing radiation in the Canadian fluoroscopy cohort study and a comparison with lung cancer mortality in the atomic bomb survivors study. *Radiat Res* 142:295–304
 79. Howe GR, McLaughlin J (1996) Breast cancer mortality between 1950 and 1987 after exposure to fractionated moderate-dose-rate ionizing radiation in the Canadian fluoroscopy cohort study and a comparison with breast cancer mortality in the atomic bomb survivors study. *Radiat Res* 145:694–707
 80. Little MP, Weiss HA, Boice JD Jr, Darby SC, Day NE, Muirhead CR (1999) Risks of in Japanese atomic bomb survivors, in women treated for cervical cancer, and in patients treated for ankylosing spondylitis. *Radiat Res* 152(3):280–292
 81. Little MP, Boice JD Jr (1999) Comparison of breast cancer incidence in the Massachusetts tuberculosis fluoroscopy cohort and in the Japanese atomic bomb survivors. *Radiat Res* 151:218–224
 82. Lubin JH, Boice JD Jr, Edling C, Hornung RW, Howe GR, Kunz E et al (1995) Lung cancer in radon-exposed miners and estimation of risk from indoor exposure. *J Natl Cancer Inst* 87(11):817–827
 83. Cardis E, Gilbert ES, Carpenter L, Howe G, Kato I, Armstrong BK, Beral V, Cowper G, Douglas A, Fix J, Fry SA, Kaldor J, Lave C, Salmon L, Smith PG, Voelz GL, Wiggs LD (1995) Effects of low doses and low dose rates of external ionizing radiation: cancer mortality among nuclear industry workers in three countries. *Radiat Res* 142:117–132
 84. Cardis E, Vrijheid M, Blettner M, Gilbert E, Hakama M, Hill C, Howe G, Kaldor J, Muirhead CR, Schubauer-Berigan M, Yoshimura T, Bermann F, Cowper G, Fix J, Hacker C, Heinmiller B, Marshall M, Thierry-Chef I, Utterback D, Ahn YO, Amoros E, Ashmore P, Auvinen A, Bae JM, Solano JB, Biau A, Combalot E, Deboodt P, Diez Sacristan A, Eklof M, Engels H, Engholm G, Gulis G, Habib R, Holan K, Hyvonen H, Kerekes A, Kurtinaitis J, Malke H, Martuzzi M, Mastauskas A, Monnet A, Moser M, Pearce MS, Richardson DB, Rodriguez-Artalejo F, Rogel A, Tardy H, Telle-Lamberton M, Turai I, Usel M, Veress K (2005) Risk of cancer after low doses of ionising radiation: retrospective cohort study in 15 countries. *Br Med J* 331(7508):77
 85. Cardis E, Vrijheid M, Blettner M, Gilbert E, Hakama M, Hill C, Howe G, Kaldor J, Muirhead CR et al (2007) The 15 country collaborative study of cancer risk among radiation workers in the nuclear industry: estimates of radiation-related cancer risks. *Radiat Res* 167(4):396–416
 86. Canadian Nuclear Safety Commission (2009) Tritium releases and dose consequences in Canada in 2006. Part of the tritium studies project INFO-0793
 87. Durham Region Health Department (2007) Radiation and health in Durham region. Durham Region Health Department, Oshawa
 88. Evrard A-S, Hemon D, Morin A, Laurier D, Tirmarche M, Backe J-C, Cartier M, Clavel J (2006) Childhood leukaemia incidence around French nuclear installations using geographic zoning based on gaseous discharge dose estimates. *Br J Cancer* 94(9):1342–1347
 89. Grosche B, Lackland D, Mohr L, Dunbar J, Nicholas J, Burkart W, Hoel D (1999) Leukaemia in the vicinity of two tritium-releasing nuclear facilities: a comparison of the Kruemmel Site, Germany, and the Savannah River Site, South Carolina, USA. *J Radiol Prot* 19:243–252
 90. Jablon S, Hrubec Z, Boice JD Jr (1991) Cancer in populations living near nuclear facilities. A survey of mortality nationwide and incidence in two states. *J Am Med Assoc* 265(11):1403–1408
 91. Sever LE, Hessol NA, Gilbert ES, McIntyre JM (1988) The prevalence at birth of congenital malformations in communities near the Hanford site. *Am J Epidemiol* 127(2):243–254
 92. Spix C, Schmiedel S, Kaatsch P, Schulze-Rath R, Blettner M (2008) Case-control study on childhood cancer in the vicinity of nuclear power plants in Germany 1980–2003. *Eur J Cancer* 44(2):275–284
 93. Kaatsch P, Spix C, Schulze-Rath R, Schmiedel S, Blettner M (2008) Leukaemia in young children in the vicinity of German nuclear power plants. *Int J Cancer* 122(4):721–726
 94. Laurier D, Hemon D, Clavel J (2008) Childhood leukaemia incidence below the age of 5 years near French nuclear power plants. *J Res Pract* 28:401–403
 95. Bithell JF, Keegan TJ, Kroll ME, Murphy FG, Vincent TJ (2008) Childhood leukaemia near British nuclear installations: methodological issues and recent results. *Radiat Prot Dosim* 132(2):191–197
 96. Grosche B (2008) The “Kinderkrebs in der umgebung von kernkraftwerken” study: results put into perspective. *Radiat Prot Dosim* 132(2):198–201
 97. SSK (German Radiological Protection Commission) (2008) Assessment of the “Epidemiological Study on Childhood Cancer in the Vicinity of Nuclear Power Plants” (KiKK Study): position of the commission on radiological protection (SSK). SSK, Bonn
 98. Beral V, Fraser P, Carpenter L, Booth M, Brown A, Rose G (1988) Mortality of employees of the atomic weapons establishment, 1951–82. *Br Med J* 297:757–770
 99. Cragle DL, Watkins JP, Robertson-DeMers K (1998) Mortality among workers at the Savannah River nuclear fuels production facility. In: ASA 1998 proceedings of the section on statistics in epidemiology. American Statistical Association, Alexandria, pp 83–87. <http://www.orau.gov/ehsd/EpiProceedings-jj-9-30-98.doc>

100. Johnson P, Atkinson WD, Nicholls JL (1999) Updated analysis of mortality in workers at UK atomic weapons establishments. In: *Proceedings of the SRP sixth international symposium: achievements & challenges: advancing radiation protection into the 21st century*, Southport
101. Zablotska LB, Ashmore JP, Howe GR (2004) Analysis of mortality among Canadian nuclear power industry workers after chronic low-dose exposure to ionizing radiation. *Radiat Res* 161:633–641
102. Schubauer-Berrigan MK, Daniels RD, Fleming DA, Markey AM, Couch JR, Ahrenholz SH, Burphy JS, Anderson JL, Tseng C-Y (2007) Risk of chronic myeloid and acute leukemia mortality after exposure to ionizing radiation among workers at four U.S. nuclear weapons facilities and a nuclear naval shipyard. *Radiat Res* 167:222–232
103. Richardson DB, Wing S (2007) Leukemia mortality among workers at the Savannah River Site. *Am J Epidemiol* 166(9):1015–1022
104. Hazelton WD, Suresh MH, Curtis SB, Zielinski JM, Ashmore PJ, Krewski D (2006) Biologically based analysis of lung cancer incidence in a large Canadian occupational cohort with low-dose ionizing radiation exposure, and comparison with Japanese atomic bomb survivors. *J Toxicol Environ Health A* 69:1013–1038
105. Laurier D, Grosche B, Hall P (2002) Risk of childhood leukemia in the vicinity of nuclear installations – findings and recent controversies. *Acta Oncol* 41(1):14–24
106. Committee Examining Radiation Risks of Internal Emitters (CERRIE) (2004) Final report [downloaded 20 October 2004]. <http://www.cerrie.org/report/>
107. Cox R, Menzel H-G, Preston J (2008) Invited editorial: internal dosimetry and tritium – the ICRP position. *J Res Pract* 28(2):131–135
108. Kacena V (1967) Chemical effects of decay of incorporated radioisotopes. In: *Proceedings of the conference on biological effects of transmutation and decay of incorporated radioisotopes*. IAEA, Vienna, pp 199–214
109. Myers DK, Johnson JR (1991) Toxicity and dosimetry of tritium: a review. Atomic Energy Control Board (now the Canadian Nuclear Safety Commission), Ottawa; Advisory Committee on Radiological Protection; INFO-0377
110. Feinendegen LE, Bond VP (1971) Transmutation versus beta irradiation in the pathological effects of tritium decay. In: Moghissi AA, Carter MW (eds) *Tritium*. Messenger Graphics, Nevada, pp 221–232
111. Little MP, Lambert BE (2008) Systematic review of experimental studies of relative biological effectiveness of tritium. *Radiat Environ Biophys* 47(1):71–93
112. Prosser JS, Lloyd DC, Edwards AA, Stather JW (1983) The induction of chromosome aberrations in human lymphocytes by exposure to tritiated water. *Radiat Prot Dosim* 4(1):21–26
113. Carsten AL, Commerford SL (1976) Dominant lethal mutations in mice resulting from chronic tritiated water (HTO) ingestion. *Radiat Res* 66:609–614
114. Russell WL, Cumming RB, Kelly EM, Phipps EL (1979) Induction of specific-locus mutations in the mouse by tritiated water. In: *Behaviour of tritium in the environment*. In: *Proceedings of a symposium on the behaviour of tritium in the environment*, 16 Oct 1978, San Francisco, IAEA-SM-232/85. IAEA, Vienna, pp 489–497
115. Lloyd DC, Purrott RJ, Dolphin GW, Bolton D, Edwards AA (1975) The relationship between chromosome aberrations and low LET radiation dose to human lymphocytes. *Int J Radiat Biol* 28:75–90
116. Chopra C, Heddle JA (1988) Cytogenetic measurements of the relative biological effectiveness of tritium. Report INFO-0287. Atomic Energy Control Board (now the Canadian Nuclear Safety Commission), Ottawa
117. Kocher DC, Apostolaei AI, Hoffman FO (2005) Radiation effectiveness factors for use in calculating probability of causation of radiogenic cancers. *Health Phys* 89(1):3–32
118. Environment Canada (EC), Health Canada (HC) (PSL2) (2003) Releases of radionuclides from nuclear facilities (impact on non-human biota). Canadian Environmental Protection Act, 1999, Priority substances list assessment report. Environment Canada, Ottawa
119. International Commission on Radiation Units and Measurements (ICRU) (1970) Linear energy transfer. ICRU Report 16, Bethesda
120. Nikjoo H, Goodhead DT (1991) Track structure analysis illustrating the prominent role of low energy electrons in radiobiological effects of low-LET radiations. *Phys Med Biol* 36:229–238
121. Chen J (2006) Radiation quality of tritium. *Radiat Prot Dosim* 122(1–4):546–548
122. Cheng YS, Snipes MB, Kropf RF, Jow HN (1995) Radiation dosimetry of metal tritides. *Health Phys* 68(Suppl 6):S53
123. Cheng YS, Snipes MB, Wang Y, Jow HN (1999) Biokinetics and dosimetry of titanium tritide particulates in the lung. *Health Phys* 76(2):120–128
124. Cheng YS, Zhou Y, Wang Y, Inkret WC, Wermer JR, Joseph R (2002) Dose estimate of inhaled hafnium tritide using the ICRP 66 lung model. *Health Phys* 82(6):817–824
125. International Commission on Radiological Protection (ICRP) (1995) Age-dependent doses to members of the public from intakes of radionuclides: part 4, inhalation dose coefficients. Publication 71, 25(3–4). Pergamon, Oxford
126. Richardson RB, Hong A (2001) Dose to lung from inhaled tritiated particles. *Health Phys* 81(3):313–324
127. International Commission on Radiological Protection (ICRP) (1989) Age-dependent doses to members of the public from intake of radionuclides: part 1. ICRP Publication 56. Ann ICRP 20(2). Pergamon, Oxford
128. International Commission on Radiological Protection (ICRP) (1993) Age-dependent doses to members of the public from intake of radionuclides: part 2. ICRP Publication 67. Ann ICRP 23(3/4). Pergamon, Oxford
129. Health and Welfare Canada (1983) Bioassay guideline 2 – guidelines for tritium bioassay, Report of the working group

- on bioassay and in vivo monitoring criteria, Environmental Health Directorate, 83-EHD-87
130. Canadian Nuclear Safety Commission (2005) Radiobioassay and dose assessment for intakes of tritium, Report of the CNSC working group on internal dosimetry, RPS-0182B, Radiation Index
 131. Atomic Energy Control Board (now the Canadian Nuclear Safety Commission) (1987) The determination of effective doses from the intake of tritiated water. Regulatory Policy Statement R-100
 132. Johnson JR (1982) The estimation of the effective dose equivalent from tritiated water exposures using tritium concentrations in urine. *Radiat Prot Dosim* 2:245–247
 133. International Commission on Radiological Protection (ICRP) (1995) Basic anatomical and physiological data for use in radiological protection: the skeleton. Publication 70, 25(2). Pergamon, Oxford
 134. International Commission on Radiological Protection (ICRP) (1979–1982) Limits for intakes of radionuclides by workers, ICRP Publication 30, Part 1 (and Supplement), Part 2 (and Supplement), Part 3 (and Supplements A and B), and Index, prepared by Committee 2, adopted by the Commission in July 1978, *Annals of the ICRP*. Pergamon, New York
 135. Trivedi A, Galeriu D, Richardson RB (1997) Dose contribution from metabolized organically bound tritium after acute tritiated water intakes in humans. *Health Phys* 73(4):579–586
 136. Trivedi A, Galeriu D, Lamothe ES (2000) Dose contribution from metabolized OBT after chronic tritiated water intakes in humans. *Health Phys* 78(1):2–7
 137. Osborne RV (1966) Absorption of tritiated water by people. *Health Phys* 12:1527–1537
 138. Peterman BF, Johnson JR, McElroy RGC (1985) HT/HTO conversion in mammals. *Fusion Technol* 8:2557–2563
 139. International Commission on Radiological Protection (ICRP) (1994) Human respiratory tract model for radiological protection. Publication 66, 24(1–3). Pergamon, Oxford
 140. Phipps AM, Kendall GW, Fell TP, Harrison JD (1990) Doses from radioactive methane. *Radiat Prot Dosim* 30:191–195
 141. International Commission on Radiological Protection (ICRP) (1996) Age-dependent doses to members of the public from intake of radionuclides: part 5. ICRP Publication 72. Ann ICRP 26(1). Pergamon, Oxford
 142. International Commission on Radiological Protection (ICRP) (2001) Doses to the embryo and fetus from intakes of radionuclides by the mother. Publication 88, 31(1–3). Pergamon, Oxford
 143. International Commission on Radiological Protection (ICRP) (2004) Doses to infants from ingestion of radionuclides in mothers' milk. ICRP Publication 95, 34(3–4). Pergamon, Oxford
 144. Butler HC, Leroy JH (1965) Observation of biological half-life of tritium. *Health Phys* 11:283–285
 145. Rudran K (1988) Significance of in vivo organic binding of tritium following intake of tritiated water. *Radiat Prot Dosim* 25(1):5–13
 146. Balonov MI, Dolgirev EI, Likhtarev IA (1974) Exchange kinetics and dosimetry of tritium oxide in man for different routes of administration. *Health Phys* 27(4):367–375
 147. Foy JM, Schneiden H (1960) Estimation of total body water (virtual tritium space) in the rat, cat, rabbit, guinea-pig and man, and of the biological half-life of tritium in man. *J Physiol* 154:169–176
 148. Wylie KF, Bigler WA, Grove GR (1963) Biological half-life of tritium. *Health Phys* 9:911–914
 149. Snyder WS, Fish BR, Bernard SR, Ford MR, Muir JR (1968) Urinary excretion of tritium following exposure of man to HTO – a two exponential model. *Phys Med Biol* 13: 547–559
 150. Sanders SM Jr, Reinig WC (1968) Assessment of tritium in man. In: *Diagnosis and treatment of deposited radionuclides*. Excerpta Medica Foundation, Amsterdam, pp 534–542
 151. Minder W (1969) Internal contamination with tritium. *Strahlentherapie* 137:700–704
 152. Lambert BE, Sharp HBA, Dawson KB (1971) An accidental intake of tritiated water. *Am Ind Hyg Assoc J* 32:682–686
 153. Moghissi AA, Carter MW, Lieberman R (1971) Long-term evaluation of the biological half-life of tritium. *Health Phys* 21:57–60
 154. Moghissi AA, Carter MW, Lieberman R (1972) Further studies on the long-term evaluation of the biological half-life of tritium. *Health Phys* 23:805–806
 155. Bennet BG (1972) The radiation dose due to acute intake of tritium in man. USAEC Report, HASL-253
 156. Trivedi A, Richardson RB, Galeriu D (1994) Dose from organically bound tritium after an acute tritiated water intake in humans. Atomic Energy Control Board (Canada) Research Report INFO-0598
 157. Taylor DM (2003) A biokinetic model for predicting the retention of ^3H in the human body after intakes of tritiated water. *Radiat Prot Dosim* 105:225–228
 158. Takeda H, Kasida Y (1979) Biological behavior of tritium after administration of tritiated water in the Rat. *J Radiat Res* 20:174–185
 159. International Commission on Radiological Protection (ICRP) (1959) Recommendations of the international commission on radiological protection: report of committee II. ICRP Publication 2. Pergamon, London
 160. Ichimasa Y, Ichimasa M, Shiba T, Oda M, Akita Y (1986) Fixation of tritium gas by rats. *Radiat Prot Dosim* 16:127–130
 161. International Commission on Radiological Protection (ICRP) (1994) Dose coefficients for intakes of radionuclides by workers. Publication 68, 24(4). Pergamon, Oxford
 162. Trivedi A, Gentner NE (1999) Dosimetry and health effects of tritium, AECL RC-2243. AECL, Chalk River
 163. Eakins JD, Hutchinson WP, Lally AE (1975) The radiological hazard from tritium sorbed on metal surfaces. *Health Phys* 28(3):213–224
 164. Johnson JR, Dunford DW (1985) Dosimetric models of ^3H from skin absorption following contact with T2-contaminated surfaces. *Health Phys* 48(1):110–113

165. Trivedi A, Cornett RJ, Galeriu D, Workman W, Brown RM (1997) Daily tritium intakes by people living near a heavy-water research reactor facility: dosimetric significance. AECL-11648, COG-96-333-I
166. Richardson RB, Trivedi A, Greenstock CL (1998) Dosimetry of organically bound tritium derived from diet – phase 1. AECB RSP-068
167. Richardson RB (2001) Dosimetry of organically bound tritium derived from diet – phase 2. Report to the Canadian nuclear safety commission, Contract no. 87055-8-5044/001/SS
168. Richardson R, Dunford DW (2003) A biochemical-based model for the dosimetry of dietary organically bound tritium. *Health Phys* 85(5):523–538
169. Richardson RB, Dunford DW, Peterson SR (2001) Influence of gender differences in the carbon pool on dose factors for intakes of tritium and ^{14}C -labelled compounds. *Health Phys* 81(3):302–312
170. Komatsu K, Okamura Y, Sakamoto K (1990) Radiation dose to mouse liver cells from ingestion of tritiated food and water. *Health Phys* 58:625–629
171. Müller WU (2008) Comment on the invited editorial 'effectiveness of tritium beta particles'. *J Res Pract* 28:249–252
172. Müller WU (2008) Reply to the reply to comment on the invited editorial 'effectiveness of tritium beta particles'. *J Res Pract* 28:423–426
173. Bridges BA (2008) Reply to comment on the invited editorial 'effectiveness of tritium beta particles'. *J Res Pract* 28:249–252
174. Bridges BA (2008) A further reply. *J Res Pract* 28:423–426

Tropical Health and Sustainability

J. KEVIN BAIRD

Eijkman–Oxford Clinical Research Unit, Jalan
Diponegoro, Jakarta, Indonesia

Centre for Tropical Medicine, Nuffield Department of
Medicine, University of Oxford, Oxford, UK

Article Outline

Glossary

Definition of the Subject and Its Importance

Introduction

Forging Tools

Providing Tools

Diagnostics Are Critical

Insecticides Are Critical

Drug Delivery

The Vaccine Development Gambit

R&D Investment

Sustainable S&T in Tropical Health

Southern Scientific and Technological Development

Future Directions

Acknowledgments

Bibliography

Glossary

ACT Artemisinin-combined therapies, a large class of paired drugs that represent the front-line treatments for malaria globally.

BCMS Board for the Coordination of Malaria Studies, a US government entity established during World War II and later dissolved to coordinate management of the development of new antimalarial therapies.

BMGF Bill & Melinda Gates Foundation.

CFR Case fatality rate, the percentage of patients with a given clinical condition not surviving.

DDT Pesticide ((1,1,1-trichloro-2,2-di(4-chlorophenyl) ethane)), once often used in IRS applications.

G6PD Glucose-6-phosphate dehydrogenase, an inherited deficiency of which causes patients to be vulnerable to mild to severe drug-induced acute intravascular hemolysis.

GFATM Global Fund for AIDS, Tuberculosis and Malaria, a consortium of International donors committing resources to those health issues.

GMEC Global Malaria Eradication Campaign of the 1950s and 1960s.

GMP Good manufacturing practice, a high standard of manufacturing certified by experts.

IRS Indoor residual spraying, in malaria control the practice of spraying the interior walls of homes with insecticide in order to attack mosquitoes that feed on humans.

ITN Insecticide-treated net, or a net that covers a bed at night to protect sleeping people from mosquito bites and malaria infection.

Malaria Infection by protozoan parasites (*Plasmodium* species) carried by mosquitoes, often serious and fatal.

MMV Medicines for Malaria Venture, a public-private partnership committed to developing and licensing new antimalarials therapies.

North/south Generally characterizes global geographic divide between so-called developed and developing nations.

R&D Research and development, deliberate, systematic application of S&T in striving toward specific understanding or objectives.

RDT Rapid diagnostic test, an immunochromatographic, point-of-care kit used at village level to diagnose pathogens (malaria) at low cost.

S&T Science and technology, techniques for expanding understanding of the physical universe.

Sustainability Technologies or systems that operate effectively in the absence of long-term external financial or technical assistance.

Tropical health Health issues specific to the tropical zones, especially endemic infectious diseases and underdeveloped healthcare delivery.

WHO World Health Organization.

WRAIR The Walter Reed Army Institute of Research in Washington, DC, where development of antimalarials by the US government has been supported.

Definition of the Subject and Its Importance

Tropical health may be considered relevant to the range of human maladies that occur either predominantly in the tropics or become exacerbated by poorly resourced healthcare delivery systems typical of many nations in the tropics. More often, it is both of those factors, as with malaria. Sustainability implies the introduction of novel interventions aimed at mitigating the burdens and risks of such maladies by improving or replacing existing instruments that may be inadequate or suboptimal. Sustainability further implies a design of those instruments amenable to the capacities of those left to use them, following the inevitable retreat of sponsors and their fiscal and technical resources. Further, the instrument must prove effective in mitigating the problem at which it aims. The research and development (R&D) of sustainable technologies against tropical infections represents a relentlessly urgent task, often linked to preventable deaths in very substantial numbers.

The limited ability of underdeveloped healthcare systems to implement many or most new health technologies greatly compounds the difficulty of achieving sustainability. New science and technology (S&T) must be adapted to settings often completely alien to its developers, both in terms of ability to satisfactorily apply it and its suitability in achieving desired impacts. The subject of tropical health and sustainability is thus

concerned with the gap between new technological tools and the ability of tropical healthcare systems to leverage them in improving human health. Further, the intended tools may be poorly conceived as a consequence of fundamental misunderstanding of the problem and, even with appropriate implementation, contribute little to real progress against it.

These distinct gaps, such as the poor grasp of the health problem on one side and of new technologies on the other, may be historically viewed along a North/South geographic divide. The nations that industrialized in the nineteenth century and led the scientific and technological revolution of the twentieth century were predominantly above the Tropic of Cancer. Below that divide, most nations remained largely agrarian and relatively less developed as far as S&T were concerned. Today, many nations of that South are often characterized as developing or resource limited and nations of the North, of course also continuously developing, are characterized as developed or resource rich. The North still commands global scientific and technological superiority, at least within the context of how the North views that critical field of human endeavor. Sustainable S&T in tropical health thus typically refer to tools developed in the North but put to practical and effective use in the South.

The importance of tropical health and sustainability lies in both the Northern and Southern perspectives of what this subject may mean in practice. In the North it implies applying a formidable suite of S&T capacities in improving Southern health. The Southern perspective may focus upon broad improvement of healthcare delivery by any means and not necessarily by the application of what may be poorly understood or even unknown Northern technologies. This perspective may be misconstrued in the North as the larger part of the problem, that is, a lack of awareness of the possibilities of Northern technology. This in turn leads the North to assume leadership roles in applying their R&D capacities to Southern health problems. Poor grasp of those health challenges may instead be the larger part of S&T sustainability failures in addressing them. This entry explores that problematic paradigm and alternatives to it.

Introduction

The frontiers of science and technology (S&T) represent increasingly expensive investments in a future

driven by expanding complexity and sophistication of research and development (R&D) systems. Nations endowed with the fiscal and technical wherewithal for such investments have historically dominated important scientific and technological progress. Moreover, these advances have often been driven by conspicuous national self-interest, for example, geopolitical competition with the Soviet Union catalyzed the US National Aeronautics & Space Administration into a technology behemoth. The US Department of Defense may be described in the same terms. Quite distinct from military superiority, nations also recognize the extraordinary economic potential of S&T, and it is increasingly linked to the political and economic security of nations. Contemporary reviews of national S&T agendas express this explicitly [1, 2]. The advancement of other nations in S&T may be construed as a threat by the empowering of potential economic competitors or even political or military rivals. One nation or group of nations may be inherently reluctant to promote the advancement of the S&T capacities of other nations.

Northern nations certainly acknowledge the vital importance of improving the global human condition and they do not lack in humanitarian compassion and generosity. Vast resources and technical know-how have flowed from North to South in the hope of alleviating endemic human suffering in much of the tropics. According to the Organization for Economic Development and Assistance, \$104 billion was donated to developing countries by developed countries in 2007 (<http://www.oecd.org/dataoecd/47/25/41724314.pdf>). About \$6 billion of that figure went to the health sector (in 2006), and recipient governments expended three times that amount from their own coffers for the same purpose [3]. However, very few of these resources are applied to advancing Southern or Northern health S&T/R&D per se – these funds almost invariably go to the application of proven rather than experimental technological solutions to human health problems. The North aims to apply proven technologies against Southern health problems through the broad aid schema. These funds and their application in the business of improving Southern health may be thought of as downstream from the S&T/R&D capacities aimed at improving the efficacy of these interventions.

In the S&T context of this encyclopedia, focus may be given to those R&D capacities and their effectiveness

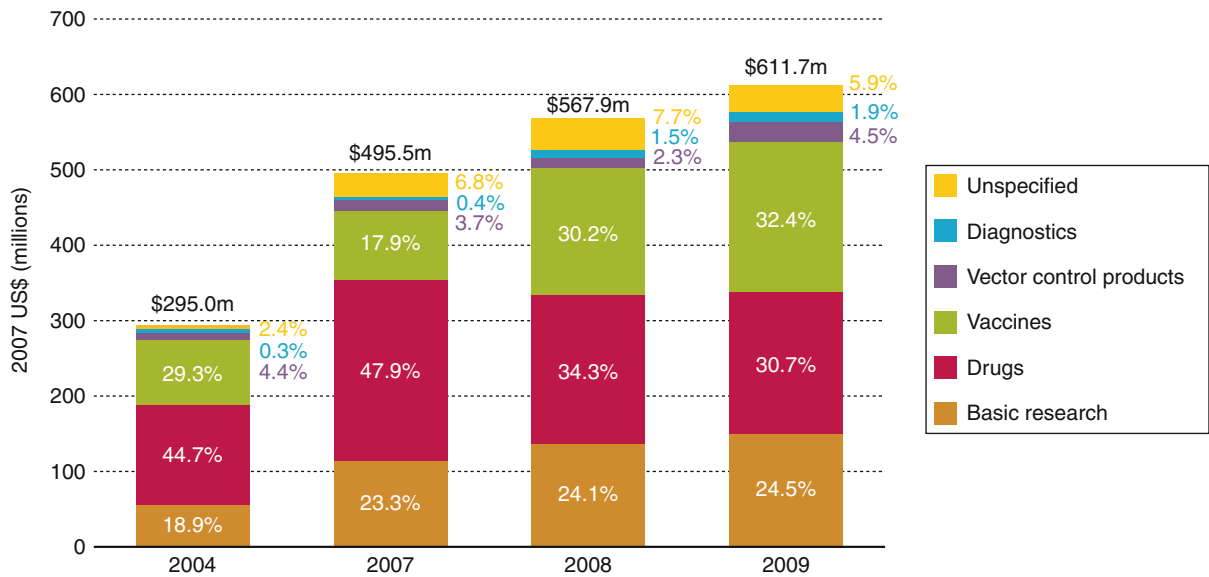
in real support of the downstream application of interventions against human health problems. This entry further focuses on a single important health issue, malaria, as an example of how Northern S&T works on a serious Southern health problem. The relatively vast sums available for malaria R&D (a recent development) likely create dynamics distinct from neglected tropical diseases like leishmaniasis, trypanosomiasis, filariasis, and many others. However, “neglected” may remain an appropriate term for malaria despite the funds being expended on it. An important question is the efficacy or impact of those funds in creating technologies that contribute to solving the problem.

Malaria causes very high burdens of morbidity and mortality, in addition to significantly stunting economic development [4]. At least hundreds of thousands perish each year and hundreds of millions suffer debilitating illness [5, 6]. In the tropical South, such burdens have remained entrenched despite the emergence of modern medications against the parasite and insecticides against the mosquito vectors over 90 years ago [7]. Malaria is preventable, treatable, and its cycle of transmission may be permanently broken with available S&T. This is known with certainty. Why then does malaria remain such an enormous human problem? This question strikes at core key issues in tropical health sustainability in the context of North-South R&D collaboration.

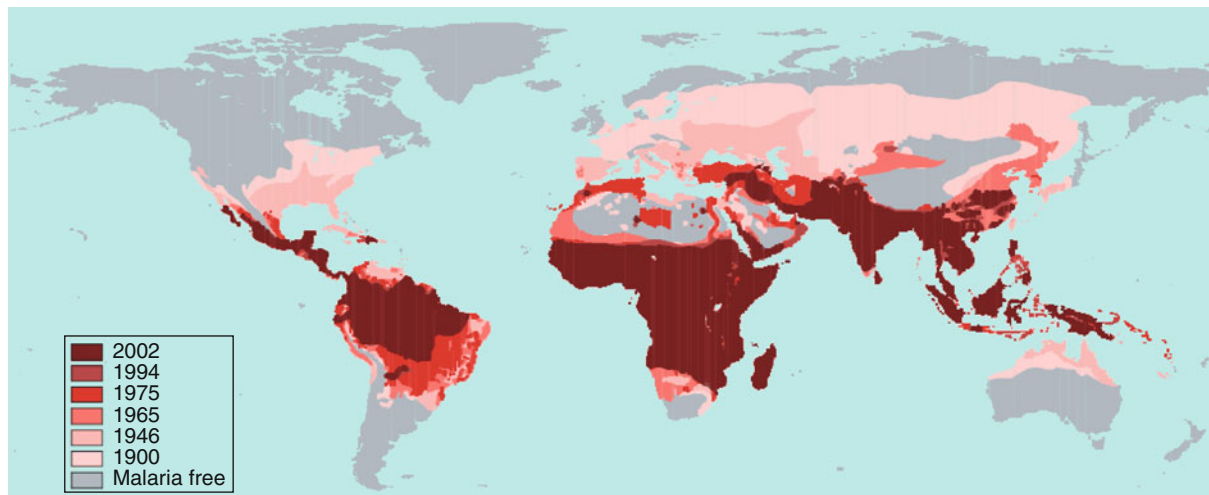
In 2011, a consortium of organizations funded by the Bill & Melinda Gates Foundation (BMGF) published an enormously informative and useful summary of global malaria R&D activities and financing (Fig. 1), predominantly for the period 2004–2009 [8]. The perspective taken in this entry reaches further back, to the Northern commitment to and later abandonment of malaria as an S&T endeavor between 1900 and 1965. The modern era of malaria R&D began around 1980 when Northern laboratories began applying fast-breaking biotechnological methodologies in earnest. These broader views provide insights on North-South S&T dynamics and a means of assessing the strengths and weaknesses historically embedded within them.

Forging Tools

The analogy of endemic malaria as a craftsman’s problem to be worked by applying the range of tools in



Tropical Health and Sustainability. Figure 1
Global malaria R&D spending (Taken from reference [8] with permission of the Malaria Vaccine Initiative)



Tropical Health and Sustainability. Figure 2
The retreat of malaria to current endemic bounds since 1900 (Taken from reference [9] with permission of the authors and publishers)

a toolbox provides a useful conceptualization of human management of that problem. The craftsman's finished project may be thought of as a void once occupied with the human misery caused by malaria. The analogy finds further usefulness in recognizing the craftsman as representing the governments and institutions in endemic zones doing the physical work of crafting

that project. They pick up the tools and use them. In the first half of the twentieth century, the Northern craftsman designed and forged the tools in his toolbox. As may be seen by the retreat of malaria between 1900 and 1965 [9] (Fig. 2), this approach produced effective outcomes. Malaria disappeared from the North and began to be thought of as an exclusively Southern

disease. The retreat of malaria involved complex dynamics [7], especially economic, but its initiation only after science generated sufficient understanding and tools makes compelling argument for the vital efficacy of S&T in achieving such gains.

A glance at where malaria remains entrenched today reveals where that early Northern toolbox has proved inadequate to the task of elimination. Experts will see determinants of that failure according to their respective area of expertise. The biologist will point to the tropical climate as especially conducive to the transmission of malaria parasites by their anopheline mosquito definitive hosts. The medic will highlight inadequate healthcare systems. The economist will blame the relatively impoverished condition of the nations composing the malaria belt. Sociologists and politicians will note armed conflict and poor or unstable systems of governance dotted across that belt. And each expert would be partly correct. All of these issues contribute to the contemporary intransigence of endemic malaria in the tropics and even some temperate zones, despite the long availability of technologies adequate to the problem elsewhere. Northern R&D strives to outfit the Southern craftsman with an improved toolbox that may overcome these factors.

The Northern R&D community engaging malaria sees itself as the designer and forger of those tools. Historic Northern dominance in S&T and inarguably superior standing in R&D capacities relative to those of the South, logically lead to the assumption of this role. Northern laboratories engaging the malaria problem express this role explicitly. As one example among a wide range of Northern R&D institutions, the following comes from the US National Institutes of Health, National Institute for Allergy & Infectious Diseases (NIAID) 2008 publication, “NIAID Research Agenda for Malaria,” and the section headed, “Malaria Prevention & Control Strategies” [10]: “*To reverse the trend toward rising malaria prevalence and its return to countries where it had been eliminated, new drugs, diagnostics, and vector management tools, as well as effective vaccines, are urgently needed. There also is an acute need to understand the factors that favor the effective use of new interventions and of prevention and control interventions already available. To develop both new products and a more complete understanding of factors that assure their effective introduction and use, public-private*

partnerships are essential.” The message also comes across in the 2011 “Staying the Course?” monograph [8]: “... *investment in malaria R&D would create new tools that would more rapidly decrease the malaria burden...*”

The presumed superior usefulness of those Northern tools to the task at hand, as opposed to those that could possibly be developed by the Southern craftsman himself, may be rationally challenged. However, Southern R&D capacity today cannot match that in the North and long-term investment in Southern S&T as an alternative approach has been, and continues to be, trumped by the often expressed sentiment, most recently in the malaria R&D monograph [8], that: “*Providing slow and inadequate funding over the next five to six years will not only delay the time when donors can begin to reduce their malaria R&D funding but – more importantly – will unnecessarily delay saving millions of lives in the developing world.*” The unspoken corollary is the humanitarian urgency in fighting malaria renders the long and difficult work of building Southern R&D capacity strategically irrelevant. Northern R&D strategizes a solution for global malaria upon what is effectively faith in its ability to deliver definitive solutions in the short term.

The contemporary position of an organization called the Multilateral Initiative on Malaria (MIM) supports this assessment. MIM was formed for the specific purpose of creating malaria R&D capacities in Africa. The malaria R&D monograph [8] devotes a single sentence to the organization and offers no insight on its funding levels, despite the rich details on funding embedded throughout the book. MIM’s funding page on its website (www.mimalaria.org) was blank in July 2011. It may be surmised from other numbers provided in the monograph that the MIM budget was perhaps about \$1.4 million dollars in 2009 (it is also unclear where that funding originated) [8]. It may thus be seen that of the \$612 million dollars allocated to malaria R&D in 2009, 0.2% of that was committed to building Southern malaria R&D capacity, and only in Africa. At least within the Northern R&D agenda expressed in that monograph, standing Northern R&D capacity and the urgency of lives lost seems to preclude meaningful investment in long-term transfer of R&D capacities southward.

Providing Tools

If Northern R&D is to be the designated driver of tool design and delivery to the Southern craftsman, an assessment of its performance in this role may be instructive and useful. Performance evaluations look backward rather than forward, even though ultimately aimed at better performance in the future. Assessments of possible or hoped-for performance add little to the utility of such evaluations. Useful evaluation of R&D performance thus demands suspension of its primary focus, the future. How has the Northern R&D posture of the past delivered upon a future represented by today? What lessons may be gathered from that experience and applied in a strategic sense to better performance in reaching a planned and sustainable future?

The primary tools of malaria control today are rapid diagnostic tests (RDTs), artemisinin-combined therapies (ACTs), insecticide-treated bed nets (ITNs), and, increasingly, indoor residual insecticide spraying (IRS) [11]. None of these tools emerged from mainstream contemporary Northern malaria R&D imperatives or programs (Table 1). RDTs and ITNs were almost entirely commercially developed, and IRS represents an almost abandoned and recently rediscovered tool developed almost a century ago. The vital artemisinin component of ACTs came from the Chinese Army, and the concept and promise of ACTs emerged in the late 1980s and early 1990s from individuals (particularly Prof. Nicholas White at Mahidol University in Thailand) and private industry

(Novartis began codeveloping an ACT with public Chinese partners in 1990) rather than any Northern R&D agenda impelling them in that direction [12]. Up to the present day, with the exception of drugs, the other three vital tools in the standing malaria control toolbox represent the two most severely underfunded avenues of the malaria R&D agenda: diagnostics (1% of funding) and mosquito vector-based interventions (4% of funding) [8].

The proposed strategy to deal with this underfunding is to hold all other areas (basic, drugs, vaccines) stable and to reserve the hoped-for 2% increases in annual funding to those neglected avenues [8]. In short, the expressed strategy is to give those two areas, currently funded at \$30 million, an incremental (and hoped-for) increase of about \$12 million per year. If this report represents Northern S&T consensus (it does not purport to do so), nothing in basic, drugs, or vaccines research may be sacrificed beyond the savings of achieved targets (and those go to offer relief to donors). The Northern toolmaker's investment preferences seem to reflect optimism for delivery of more practical and useful tools than diagnosis and warding off mosquitoes. Drugs and vaccines (and the basic research that ultimately drives those two avenues) command the Northern R&D agenda at 89% of expenditures between 2004 and 2009 [8]. Has the toolmaker gotten this balance right?

Another legitimate question may be, "Who is this toolmaker?" There is no formal body that gathers and determines the direction and magnitude of malaria

Tropical Health and Sustainability. Table 1 The malaria control toolbox 2011

Tools	State of the art	Era developed	Sector developed	Utility
<i>Diagnosis of infection</i>	Immunochromatographic cassettes	1990s	Private	Limited
<i>Diagnosis of G6PD deficiency</i>	NADP + reduction assays	1960s	Private	Poor
<i>Therapy against acute attack</i>	ACTs	1990s	Private	Good
<i>Therapy against relapse</i>	Primaquine	1940s	US Army	Poor
<i>Therapy against transmission</i>	Primaquine	1940s	US Army	Poor
<i>Medical insecticides</i>	Pyrethroids, carbamates, organophosphates	1970s	Private	Poor
<i>Insecticide treated nets</i>	Long-lasting pyrethroid impregnated nets	1990s	Private	Limited

R&D funding decisions. The malaria R&D monograph [8] is simply a summary of where investments have been made and by whom. The data are presented for the expressed purpose of guiding those bringing fiscal resources to the malaria R&D table. Thus, the people embedded in funding organizations and having the responsibility of distributing those funds constitute the toolmaker and his collective resources – money, ideas, expertise, a plan, and the will to achieve it.

In 2009, \$612 million was expended on malaria R&D. The \$99 million from this figure expended by biotechnological firms may be discounted because the funds do not set anyone else's research agenda – they overwhelmingly fund their own R&D imperatives and operations [8]. Two thirds of the remaining \$513 million came from the BMGF (\$184 million), the US NIH (\$116 million), and the US Department of Defense (\$38 million). Three European donors followed with a collective \$77 million (European Commission, Wellcome Trust, UK Medical Research Council). The balance of \$98 million came from a variety of other donors. Decision makers in R&D funding among these organizations, collectively and proportionally, embody the major global malaria toolmaker.

Diagnostics Are Critical

Malaria in the real world is poorly understood by most in the North. The terminology “real world” captures the sum of complex medical, scientific, ecological, social, economic, and political determinants of endemic transmission and disease. It also expresses real (evidence-based) versus imagined (supposition-based) understanding of the character of entrenched endemic malaria. Strategists look at malaria in terms of the burdens of morbidity and mortality imposed by each of the five species of plasmodia routinely infecting humans. This entry considers only two of those species, *Plasmodium falciparum* (cause of falciparum malaria) and *Plasmodium vivax* (cause of vivax malaria) because these likely represent the vast majority of the global malaria burden.

Estimating Burden of Morbidity

Few in the North appreciate the great distance between the real malaria case numbers and those reported by

Ministries of Health (MoH) and globally summarized by the World Health Organization (WHO) each year [13]. The crucial importance of that distance is even less appreciated as a key determinant of S&T sustainability failures against endemic malaria. It reflects not just incomplete understanding of the magnitude of those burdens [6, 14], but also what ultimately drives the unseen and unknown burdens in place and time, that is, in the real world. This vulnerability is driven by unsustainable S&T for the diagnosis of malaria in endemic zones.

Data from Indonesia, to be considered typical among endemic nations, illustrate unsustainable diagnostic S&T and the consequences of the failure to mobilize a nearly adequate R&D response to it. In 2008, WHO reported between 1.4 and 1.8 million annual malaria cases in Indonesia during the years 2000–2006. In 2009, they modified estimates for the same years to between 1.3 and 3.0 million. These were based on confirmed cases reported to them by the Indonesian authorities ranging from 0.22 to 0.44 million per year [13, 15]. The differences in those numbers reflect suppositions regarding the efficiency of Indonesian diagnosis and reporting systems. Studies by the Malaria Atlas Project (MAP) based upon prevalence surveys and Bayesian statistical modeling, estimated 12 million clinical attacks by falciparum malaria alone [16]. Capture of laboratory-confirmed and reported cases of malaria reflected by the numbers provided by MoH implies a roughly 10% case diagnosis rate by the WHO estimates of total cases. That rate drops to less than 1% using the MAP case estimate which excludes vivax malaria. By either estimate, despite the order of magnitude difference between them, the majority of malaria cases in Indonesia go undiagnosed and unreported [17].

The Ministry of Health of Indonesia, like almost all others, reserves therapy with ACTs to cases having a confirmed diagnosis, and all others receive ineffective chloroquine or sulfadoxine/pyrimethamine because there are few practical options [18]. They do so under the rules of the Global Fund for AIDS, Tuberculosis, and Malaria (GFATM) – the donor paying the bill for most ACTs in Indonesia. GFATM reasonably does not wish to have precious ACTs wasted for any given febrile illness. It may be appreciated that this enormous problem is not a challenge of better drugs, but better

diagnosis. Failure to deliver diagnostic services and reporting directly causes most malaria in endemic zones to go undiagnosed, unreported, and not treated.

Estimating Burden of Mortality

Reports and estimates of malaria mortality rates in Indonesia also range freely across several decimal places. The Government of Indonesia's Central Bureau of Statistics conducted national Household Health Surveys in 1995 and 2001, and estimated 30,000–38,000 deaths due to malaria had occurred in each of those years [17]. In 2001, the WHO reported 68 deaths caused by malaria in Indonesia [15]. This reported number was 494 in 2006, and WHO estimated the number to be closer to 3,000 deaths [15]. A recent study in India challenged the government's estimates of malaria mortality at 15,000 deaths in 2008 [19]. Investigators conducted verbal autopsies (interviewing surviving family members of people lost to premature death on the clinical nature of the fatal illness), using a rigorous sampling methodology to represent the broader population. They analyzed 122,000 deaths during 2001–2003, and their findings led to an estimate of 205,000 deaths annually from malaria, and death by malaria carried a 1.8% risk by age 70. They found that 86% of deaths caused by malaria did not occur in any healthcare facility, and thus went unreported. Verbal autopsy methodologies of course come with uncertainties, but it may be that fewer than 1 in 10–100 or more deaths caused by malaria are reported from endemic zones. Unsustainable S&T for diagnosis and reporting explain these dangerous gaps in understanding mortality burdens and their demographic/geographic distributions.

The species of plasmodia responsible for death cannot be ascertained in verbal autopsy, but it is now known that in India specifically, *P. vivax* at least occasionally and perhaps routinely causes death in malaria patients [20–23]. The same has been found in other regions [24–31]. Long known as “benign tertian malaria,” such reports of fatal vivax malaria have been met with appropriate, albeit stubborn skepticism in malariology circles. The conventional view has been that only the “malignant tertian malaria” of *P. falciparum* routinely causes the range of life-threatening syndromes linked to malaria.

Benign Tertian Malaria Fallacy

Careful study of the genesis of the term “benign tertian malaria” in the late 1800s and the clinical experience with malaria therapy of neurosyphilis in the 1920s and 1930s provides important insights on the contemporary view of vivax malaria as an insignificant contributor to global malaria mortality. The notion of benign and malignant malarias had been well known long before Laveran's discovery of the plasmodia in 1880 [32]. After broad acceptance of species identity between *P. falciparum* and *P. vivax* at around 1890, a few elite clinical study centers in Europe and America sought to reconcile taxonomic identity to those clinical classifications of malaria. As pointed out by Kitchen in 1949 [33], taxonomic identity should have been the frame of reference rather than ambiguous and confused clinical classifications. As early as 1901, a pathologist in New York City describing autopsy findings of a fatal case of vivax malaria challenged the certainty of *P. vivax* as a benign infection [34]. He wrote, “*While the statement of the Italian authorities has long held true that no autopsy in a case of malaria with infection by the large tertian parasite, as the infection is never fatal, the present case requires modification of that view. ...*”

The advent of malaria for therapy of neurosyphilis in the 1920s added substantial depth to the understanding of clinical course and taxonomic identity. Despite contemporary views by some that vivax malaria in those patients did not cause severe or threatening disease [35], death caused by vivax malaria carried consistent and apparently strain-specific risks ranging from 3% to 15% case fatality rates (CFR). At treatment centers across the USA, the CFR was extraordinarily consistent at 3–6% after therapy had been optimized [36–41]. Prior to both learning to screen out poor candidates for treatment and gaining expertise in successful clinical management, the CFR had been between 15% and 30%. At least one series of fatal treatments had been followed by autopsy studies that confirmed malaria as the cause of death (rather than any underlying disease, including neurosyphilis) in at least half of the patients [40]. Most physicians reporting these experiences expressed the conviction that malaria had killed most of their nonsurviving patients. Nicol [42] wrote in 1932, “*One must acknowledge that, though in a few cases death is caused by*

intercurrent disease, malaria, if not directly responsible, must be regarded as a contributing factor; in most cases the malaria itself is the cause." It should be recalled that in that era unsuccessful therapy of neurosyphilis would almost always end in death for the patient. This fact, and the only 35% efficacy of malaria therapy against the disease under the best of circumstances [42], explains the aggressive and extreme treatment. More severe and punishing repeated malaria paroxysms (episodic extreme chills and spiking fevers) improved the odds of successful therapy. Nicol [42] closed the above-cited quotation with, "*Risks, however, must be taken in treating a disease which, if untreated, proves fatal.*"

Clinics in the UK usually used the Madagascar strain of *P. vivax* and this strain for treatment of neurosyphilis typically killed 10–14% of patients [43]. Clinics in the Netherlands used a local strain of *P. vivax* that carried a 0% CFR, but had very low therapeutic efficacy against neurosyphilis. A switch to the Madagascar strain improved efficacy but came with an 8% CFR [44]. One clinic in the UK evaluated *P. falciparum* for therapy and reported a 4% CFR [45]. It may thus be appreciated that most strains of vivax malaria, permitted to follow a severe course of infection, may threaten life in at least 5% of patients. Contemporary studies in hospitals report CFR among patients classified as having severe illness at between 5% and 10%, a rate roughly equal to that among patients with severe falciparum malaria in the same hospitals.

The chapter on clinical vivax malaria penned by S.F. Kitchen [33] in the classic 1949 text on malaria does not deal directly with the neurosyphilis mortality data. The author describes some controversy on the issue of severe and fatal vivax malaria but dismisses it as implausible largely on the basis of this parasite rarely causing hyperparasitemia – a well-known risk factor for death in falciparum malaria and the basis of its "malignant tertian malaria" moniker. Relatively low parasitemia is routinely documented in contemporary reports of severe and fatal disease in vivax malaria, and this may reflect an unknown pathogenesis distinct from falciparum malaria. The seemingly nonthreatening low parasitemias of vivax malaria may have deceived earlier workers. Indeed, Kitchen [33] wrote, "...present knowledge concerning the provocative

potentialities of low-grade parasitemias does not permit denial of the possibility that the presence of the Plasmodium [vivax], even in very small numbers, may serve as an incitant of entirely foreign conditions. . ." Today it is known, for example, that *P. vivax* is physiologically capable of accessing the bone marrow [46] and is sometimes found in that tissue [47].

Although further studies are required to gauge the burden of severe morbidity and mortality of vivax relative to falciparum malaria in endemic zones, the infection is certainly not benign. Regardless of relative burden, the persistent view of vivax malaria as a benign entity may be considered incompatible with available evidence and probably dangerous to patients and populations living with this threat.

Vivax Malaria and Global Malaria Burden

The long-engrained notion that only *P. falciparum* kills profoundly shapes how the global malaria problem is perceived as a whole and, ultimately, strategized to deal with. A statement very similar to this may be found in popular, peer-reviewed, and authoritative technical media [48–50], "80% of cases and 90% of the world's deaths caused by malaria occur in Africa." Vivax malaria is relatively rare in sub-Saharan Africa as a consequence of the almost universal inherited absence of a molecule called Duffy factor on the surface of red blood cells which *P. vivax* requires for invasion. Thus, the dominance of *P. falciparum* on that continent, and the relative dominance of *P. vivax* elsewhere (along with the presumption of its benign nature) largely accounts for that popular supposition. As the severe morbidity in vivax malaria and very high estimated all malaria mortality data from India and elsewhere suggest [19–31], the assertion is scientifically unsound and perhaps likely to be a gross distortion of malaria as it occurs in the real world.

The statement likely overstates the relative burden of falciparum malaria and simply dismisses the contribution of vivax malaria to the global burden of morbidity and mortality. Hay and colleagues [16] estimated African clinical attacks of falciparum malaria accounted for 60% of the global burden of that species. Almost all of the remaining estimated attacks by falciparum malaria occurred in Central, South, and Southeast Asia, where the substantial burden of vivax

malaria, yet to be reliably estimated, also occurs [51]. The number of clinical attacks caused by *P. vivax* is estimated between 100 and 400 million [9]. As has been explained, the mortality burden imposed by this parasite (and very likely by *P. falciparum* as well) may be described as unknown. Adding the lower end of the vivax malaria morbidity range to the global falciparum clinical attack numbers for 2007 of Hay et al. [16] suggests at least 50% of the global burden of clinical attacks by any species occurs in Central, South, and Southeast Asia.

If careful study confirms heavy burdens of morbidity and mortality caused by endemic vivax malaria in Asia and elsewhere, the response to this threat would largely be limited to chemotherapies no longer likely to work (chloroquine), another that never worked effectively in the first place (primaquine), or others untried against this species. This hypothetical scenario contains an uncontestable core fact: the burden of morbidity or mortality of vivax malaria anywhere it occurs is not known. And yet, Northern science up to 2011 remains almost wholly focused on falciparum malaria, especially in an African context. Research on *P. vivax* between 2007 and 2009 accounted for only 3.1% of R&D funding [8]. There may be serious consequences to this possible miscalculation driven by unsustainable S&T for the diagnosis of malaria in endemic zones.

Proven resistance to the first-line treatment of acute vivax malaria, chloroquine, now occurs across the Indonesian archipelago, at sites through the Mekong region, and cases have been reported from the Indian subcontinent [52]. Another drug, primaquine, must also be administered to prevent further attacks caused by dormant forms in the liver. The efficacy of that drug has not been unambiguously measured since the 1950s. The hemolytic toxicity of primaquine in glucose-6-phosphate dehydrogenase (G6PD) deficient patients – so prevalent in endemic zones (5–25%) – coupled with the failure to field sustainable diagnostic S&T for that disorder, likely drives therapeutic effectiveness of primaquine close to null in endemic zones [53, 54]. There are no standardized means of assessing the efficacy of these drugs against this parasite, and no mechanisms of activity against the parasite or of resistance to the drugs are known. Worse still, the standing drug development paradigm, designed for and largely

serving the fielding of drugs against the acute attack of falciparum malaria, is very poorly suited for fielding new therapies for radical cure of vivax malaria [55].

RDTs Have Limited Utility

Several dozen commercially available immuno-chromatographic RDTs for malaria compete for the large volume of global sales. Most of these kits perform very well in specific settings, but in others they completely fail to meet operational requirements. Ideally, the kit is used for the diagnosis of malaria in acutely ill patients in endemic zones lacking access to a definitive microscopic diagnosis of malaria by Giemsa-stained blood film. When falciparum malaria is the offending agent and parasitemia is higher than 200 parasites/ μ L, sensitivity and specificity each come in at well above 90% in most kits. Below that threshold, sensitivity falls precipitously. In vivax malaria, the corresponding threshold is about 500 parasites/ μ L for most kits [56]. The threshold of parasitemia causing illness in nonimmune patients falls far below both of those diagnostic thresholds for RDTs [57]. Also, cross-sectional surveys in endemic zones typically show median parasitemia levels at or below those thresholds, especially for vivax malaria [58, 59]. Thus, RDTs are not suitable for diagnosis in nonimmune patients or for detection of parasitemia among asymptomatic residents of endemic zones.

Asymptomatic carriers of malaria represent the rule rather than the exception in almost all endemic zones. These infections almost universally present very low levels of parasitemia far beyond the diagnostic reach of RDTs [59]. Cross-sectional surveys in Cambodia employing ultrasensitive polymerase chain reaction (PCR) diagnostics revealed that only about half of those positives were also positive by expert microscopy, and far fewer by RDT. Thus, elimination strategies must consider mass administration of drugs in populations having very low levels of endemic malaria, as a consequence of diagnostics, inadequate to the task of finding and treating the few infected people [60]. Current RDT technology leaves wide diagnostic gaps in endemic zones.

Diagnostics R&D Performance Evaluation

The failure to develop sustainable diagnostic and reporting S&T for endemic zones may have distorted

the view of global malaria with possibly serious consequences. By the nearly complete neglect of unrecognized heavily burdened regions [61] and a heavily burdening species [62, 63], one may have been lured, by diagnostic blindness, into a position of profound strategic weakness. Assuming that vivax malaria indeed kills with some regularity in India and elsewhere, it may be reasonable to assert that one does not have even the barest grasp of malaria in the real world, that is, it is not known how many acute attacks kill how many people, where, or by which parasite, much less who is most vulnerable and why. Poor diagnostics is the reason and the Northern R&D toolmaker appears unable to prioritize that task (1% of R&D funding [8]).

Insecticides Are Critical

Since the discovery in 1880 by Alphonse Laveran that plasmodia cause malaria, and of their transmission by anopheline mosquitoes by Ronald Ross in 1898, malariologists have fiercely argued strategies of control focused upon treatment versus interventions aimed at the anopheline vectors. In the early 1900s, Robert Koch championed the former, and Ross the latter [64]. Both applied their concepts in the field (Koch with quinine in New Guinea, and Ross with crude anti-mosquito measures in West Africa) with almost no success. However, key successes in the era that followed, along with the discovery of dichlorodiphenyltrichloroethane (DDT), ensured the dominance of mosquito vector control in the global malaria eradication effort undertaken during the 1950s and 1960s and, later, misplaced disillusionment with that strategy.

Abandonment of a Proven Tool

Between 1904 and 1914, William Crawford Gorgas effectively attacked the mosquito vectors of the yellow fever and malaria in Panama that had decisively dashed the French efforts of the 1880s to build a canal [65]. Watson in British Malaya (today Malaysia) tamed malaria on economically important plantations by deliberate and highly specific environmental modifications aimed at the behavior of the local anopheline vector. His approach, called species sanitation, was successfully applied by Dutch and Indonesian malariologists in the Netherlands East Indies of the

1920s and 1930s [66]. In the same period, William Soper successfully eliminated from Brazil the imported and terribly efficient African mosquito vector of malaria, *Anopheles gambiae* [67]. Mueller discovered DDT in Germany at about the same time, and the US military brought it to bear as a defensive weapon against malaria and other insect vector-borne diseases during and immediately after the Second World War.

The technological stage was thus set in the late 1940s for the conception and strategic design of an ambitious effort to eradicate malaria through the 1950s and 1960s. The USA pushed and largely financed the WHO-led Global Malaria Eradication Campaign (GMEC), which was essentially an attack upon anophelines using principally DDT. The role of another product liberated from wartime Germany, the extraordinarily safe and effective treatment for acute malaria called chloroquine [68], was secondary or a mop-up in that strategy.

DDT exploited a behavior of anophelines – most species go indoors at night, feed upon sleeping people and, heavily laden with a blood meal, rest upon interior walls. DDT applied to those walls provided an opportunity for lethal contact with insecticide even many months after its application. Indoor residual insecticide spraying, or IRS, proved remarkably effective in malaria control in almost every setting. Global malaria numbers in the era of its systematic application plummeted, except in most of Africa where IRS was not applied in earnest for want of the logistical infrastructure to do so. In India, for example, over 800,000 deaths due to malaria each year fell to, reportedly, none [69]. Malaria transmission ceased in most of the Caribbean and was sharply curtailed all across Central and South America. Endemic malaria disappeared from Southern Europe, North Africa, Japan, Taiwan, the Korean peninsula, and most of Java [70].

Northern agriculture effectively derailed the success of DDT by hijacking the medically important product. The aerial spraying of DDT across many millions of square miles of croplands resulted in vast quantities of this very stable compound being introduced into the environment at disastrous cost. In 1971 in the USA alone, the year before such practice was legally banned, 14 million pounds of DDT was used in the agricultural sector [71]. The practice harmed beneficial insect populations and accumulated dangerously in higher

predators, even those above the Arctic Circle. Theft of a medically important instrument by Northern agriculture spilled southward. The onset of resistance to DDT by anopheline mosquitoes may have been driven by agricultural applications. In the late 1960s, as many of these problems were coming fully to light, the GMEC formally collapsed [72].

Adopting Impractical Tools

Northern agricultural insecticide industries turned to their S&T engines to solve the problem of the loss of DDT with viable alternatives. They required insecticides only stable long enough for the quick work of killing crop pests. Three classes of insecticides emerged to dominate that market: organophosphates, carbamates, and pyrethroids. These short-lived compounds, their costs, along with relatively high toxicity or noxious character make them poorly suited for IRS. Nonetheless, in the decades that followed, and up to the present day, Southern IRS against malaria and other vector-borne diseases of the tropics has depended upon these insecticides developed and licensed for Northern agricultural practices [73]. The impracticality of the new insecticides contributed to the broad contraction or collapse of IRS programs in the endemic tropics [74]. Those products stand as clear examples of the sometime very poor fit of Northern S&T imperatives and products to Southern health problems.

Cost-effective, safe, and practical long-lasting insecticides sprayed inside homes were nowhere in that Northern R&D agenda for the simple reason that Northern homes were not being invaded by insects carrying life-threatening infections.

Revitalizing IRS

Roberts and colleagues [75] documented quantitative insights on the health costs incurred by Northern aversion to DDT especially, and IRS in general. A science- and public health-led campaign narrowly averted a 2001 effort to ban DDT for any use anywhere in the world [76]. The aversion to DDT, borne of its reckless agricultural application, irrationally inhibited endorsement or support to Southern IRS operations for decades. In 2007, WHO reversed its long-standing recommendation against IRS in areas of stable transmission and today recommends DDT for such

operations [77]. In that reversal, WHO expressed, “*DDT is still needed and used for vector control simply because there is no alternative of both equivalent efficacy and operational feasibility...*” The reason for that simple truth is clear: Northern insecticides R&D abandoned the task of medical insecticides 60 years ago and there has been no Southern R&D capacity to pick it up.

Northern sociopolitical aversion to IRS has at long last diminished to greater support and emphasis on this vital tool in the control toolbox. In late 2009, for example, USAID advertised a US\$130 million grant for support of IRS operations in Tanzania. However, unleashing such resources to apply poorly suited short-lived insecticides may greatly dampen efficiency and impact. A recent study described a microencapsulated organophosphate insecticide (chlorpyrifos methyl, Dow Agrosiences) with >90% activity persisting at least 9 months when used in experimental IRS [78]. This work, and the growing realization of the threat of resistance to both pyrethroids and DDT [79, 80], emphasizes the need for much more R&D on insecticides.

Reliance upon Northern agricultural S&T for medically useful insecticides should be acknowledged as at least inefficient and perhaps impractical. The discovery and development of medical insecticides should not be simply incidental to agricultural pesticides. The malaria R&D monograph [8], however, points hopefully to Northern industrial pesticide firms as a means of supplementing the pitifully small amounts of funding committed to insecticides.

Insecticides R&D Performance Evaluation

Northern R&D on discovery of medically important insecticides for use in IRS or ITNs has been nascent since the discovery of DDT in the 1930s. This failure, viewed in light of the enormous impact of DDT against malaria in the middle of the twentieth century, may be considered the most conspicuous of all Northern R&D failures on global malaria. The stubborn and irrational Northern aversion to insecticide tools for malaria control may at last be waning, but R&D investment in earnest exploration of improving these tools remain grossly inadequate. Hopefully, however, the BMGF expressed interest in spatial repellency of some compounds by making a substantial investment in a proof-of-concept trial in Indonesia in 2009 (JKB is

an investigator in that effort), and their Grand Challenges funding scheme invested in mosquito attractant/repellent molecule discovery. Agricultural insecticide developers apply algorithms of chemical lethality and exclude compounds on the basis of repellent characteristics that would diminish opportunities for lethal contact with an insecticide product. If spatial repellency proves effective in risk reduction in endemic zones, it would reopen the universe of chemical exploration possibilities with development algorithms wholly distinct from insecticides. Perhaps not coincident with its operational efficacy in the past century, DDT happens to be one of the most effective compounds known in eliciting avoidance behavior in anopheline mosquitoes [81].

Drug Delivery

Humanity always needs new and better anti-infective drugs – diverse microbial populations mount their formidable Darwinian defenses that make old drugs useless, and new infections emerge. The commercial pharmaceutical industries of the North generate streams of new therapies in a fiercely competitive market environment that sharply hones the science and technologies (and regulatory prowess) of surviving companies. The prospect of profit drives and sustains this superb and very expensive engine of biotechnological innovation. However, at least several major multinational pharmaceutical firms have stepped into the malaria R&D arena with fairly substantial investments. During 2007–2009, they brought 14% of the \$1.68 billion R&D investment (including work on vaccines), but kept this funding almost entirely internal [8]. Industry also frequently participates in antimalarial drug development through donor-funded product development partnerships (PDPs) like the Medicines for Malaria Venture (MMV) funded by BMGF, Wellcome Trust, WHO, and others. PDPs spent \$114 million donor funds on malaria R&D in 2009. Private enterprises (both major multinationals and small to medium businesses) and academic or other not-for-profit institutions divided that pie in roughly equal halves [8].

State of the Art

Understanding these investments and the dynamics driving them requires a brief look at the class of therapies called the ACTs. The artemisinin family of drugs

(artemisinin, artesunate, artemether, arte-ether, and dihydro-artemisinin), all rapidly excreted, combines with another typically slowly excreted antimalarial drug (mefloquine, piperaquine, lumefantrine, pyronaridine, or others) to both improve efficacy and protect the artemisinins against onset of parasite resistance to them [12]. ACTs today are manufactured commercially at many factories scattered across South-east and South Asia, the Middle East, and Africa. These companies produce dozens of ACTs in as many formulations and means of packaging. The artemisinins have no patent protections, nor do the majority of drugs partnered with them. These products also share other important characteristics: almost none have been produced by the standards of Good Manufacturing Practice (GMP) as applied and certified by regulators, and few have co-formulated the components, preferring instead to package them together as separate pills. Few of these manufacturers possess the resources or influence to press for the aggressive arrest and prosecution of counterfeiters of their products. Collectively, these important flaws permit patients to take medication inadequate for cure and thereby expose each therapeutic agent to real risk of onset of resistance.

The GFATM will not fund ACT products of unproven quality. Therapies must be approved by WHO, and such approval includes certified compliance to GMP [82]. In effect, the availability of ACTs to Southern governments hinges upon access to GMP product. One strategy at work with GFATM funding policy is to push inferior drugs out of the marketplace with the superior drugs simply by making them at least as affordable. Thus, GMP product emerges as a key determinant of success in this strategy. In the 13 years of its existence, MMV has fielded two such ACTs (Coartem™, Novartis; Coarsucam™, Sanofi Aventis). Two others approach registration in 2011 (Euartesim™, Sigma Tau; and Pyramax®, Shin Poong).

Even as GMP co-formulated ACTs now approach closer to broader reality, resistance to ACTs has emerged in the Mekong region [83] where commercially driven monotherapy with artesunate and counterfeiting has been rampant [84]. Expansion of this problem into the rest of South and Southeast Asia, and ultimately Africa (as occurred with previous antimalarials), would constitute a global health catastrophe because currently there are no therapeutic

replacements [85]. This problem could possibly have been averted by early use of GMP, co-formulated ACTs, and aggressive, commercially driven protection against counterfeiting. The neglect of malaria chemotherapy as a whole by Northern pharmaceutical S&T, and also by malaria institutional R&D stalwarts during the second half of the 1900s, today has the South precariously near persistent, untreatable, and unbeatable endemic malaria. How did such neglect arise? Addressing this question requires examination of the genesis and content of the chemotherapeutics toolbox of 1965.

Chemotherapeutics Accomplishment: 1900–1965

Assessment of current strategy for delivery of drugs, as summarized above, requires consideration of where Northern S&T left off when it effectively abandoned the problem of malaria chemotherapy in 1965 as a mission accomplished. Examining malaria R&D capacities and imperatives in the preceding half century provides essential understanding of that turning point.

Scientific chemotherapy of malaria emerged in the organic chemistry crucible of the German aniline dye industry in the closing years of the 1800s. Screening for activity against parasites in bird models of malaria, the German laboratories of I.G. Farben during the 1920s and 1930s discovered many active classes and compounds [86]. The two most important were the 4- and 8-aminoquinolines, and focus upon them during the frenetic war-spurred search for better therapies in the 1940s delivered what were to be considered the ultimate antimalarial drugs.

Chemotherapy of malaria in 1940–1941 remained overwhelmingly dominated by quinine. The few synthetic antimalarial drugs in the market fared poorly due to unpleasant or even dangerous side effects. The German chemotherapists had superior drugs coming and plotted to unhinge the Dutch monopoly on chemotherapy of malaria manifest by their control of 95% of the world trade in quinine [87]. The following year, the Nazis occupied Holland and the Imperial Japanese armed forces occupied Java and Sumatra, where virtually all Dutch quinine originated. Axis powers denied the Allied forces access to the only drug then widely used to treat malaria.

The Allies understood the coming warfare would be conducted where malaria posed a serious threat.

Endemic malaria loomed ominously in the Western Pacific especially, and their worst fears were soon realized: the first offensive ground combat involving American troops in that war was at Guadalcanal in the Solomon Islands in mid-1942; those troops suffered malaria attack rates of 1,700/1,000 man-years [88]. When the US Army Americal Division was evacuated from Guadalcanal to nonmalarious Fiji, their chemoprophylaxis was withdrawn and they suffered attack rates by *P. vivax* of 3,700/1,000 man-years [89].

The US Government mobilized to solve their serious malaria problem. The malariologists enlisted in the effort, and their sponsors, viewed success in their endeavors as directly impacting the likelihood and speed of military victory, especially in the Pacific [90]. The Board for the Coordination of Malaria Studies (BCMS), managed from within the US National Science Foundation, oversaw the evaluation of over 14,000 potential drugs from discovery to clinical trials in dozens of laboratories and clinics around the USA, Australia, and Britain. Their top priority was *P. vivax* because it accounted for five of every six cases among Allied troops. The Allies quickly mobilized production of a synthetic antimalarial (lifting it from the German S&T machine that generated it) called atabrine (also called mepacrine or quinacrine) and put it to work for both prophylaxis and treatment [91]. This hasty solution, however, had serious drawbacks. The skin of troops adhering to atabrine prophylaxis turned a distinct yellow hue and compliance had to be strictly enforced. More importantly, atabrine had an unexpected and dangerous interaction with the only drug then available against relapse, pamaquine.

Pamaquine was the first synthetic antimalarial licensed and marketed for therapy against malaria (as Plasmochin). This 8-aminoquinoline compound exerted excellent activity against the then poorly understood phenomenon of relapse in vivax malaria. Its activity against the sexual stages responsible for infecting mosquitoes was also known. Following its distribution in the 1920s, however, it earned the reputation of being sometimes an exceptionally dangerous drug [92]. As is well known today, the 8-aminoquinolines as a class exhibit potentially lethal toxicity in patients with an inherited deficiency of glucose-6-phosphate dehydrogenase (G6PD). Unwitting administration to many such patients caused

fatalities [92] and the drug never saw widespread use. The Allies, however, facing their serious relapse problems in the Pacific theater of World War II, mobilized pamaquine to that zone and serious toxicity problems followed [93, 94]. The BCMS issued a classified report in early 1943 demonstrating that when atabrine (but not quinine) was administered with pamaquine, plasma levels of the latter elevated tenfold. An already dangerous drug to some became a dangerous drug to all, and in mid-1943 the US Surgeon General ordered it withdrawn as therapy against relapse (although it could still be used, and was, in lower doses against transmission) [95]. The Allies were effectively forced to surrender to their serious relapse problem.

The BCMS resolved to address this problem. They screened several hundred candidate 8-aminoquinolines for toxicity in rats and monkeys and sent 24 drugs to clinical trials in prisoner volunteers in the USA. That screening had critically important limitations, but the BCMS in its haste to deliver a replacement therapy accepted those acknowledged shortcomings. The screening for antimalarial activity of those compounds in birds, for example, was understood to be nearly irrelevant and the developers largely ignored those data [90]. Efficacy could only be ascertained against clinical relapse in humans. Knowing the 8-aminoquinolines to be uniquely effective against clinical relapse as a class, and relatively toxic, the screening became a search for the least toxic 8-aminoquinoline without a means of screening for efficacy before evaluation in clinical trials. By almost any contemporary standard of science, this search for a replacement therapy for pamaquine may be characterized as incomplete and barely adequate. The product thus discovered reflected these qualities.

The superior candidate drug to emerge from this program was primaquine. It exhibited a slightly higher therapeutic index than pamaquine, and its anticipated toxicity (acute intravascular hemolysis) in subjects with G6PD deficiency, discovery of which occurred in that clinical development program, was confirmed [96]. As the Korean War unfolded during the early 1950s primaquine saw wide application with good therapeutic results. Chloroquine had been registered as therapy of choice against acute attacks of vivax malaria in 1946 [97]. These two companion drugs thus became therapy of choice for radical cure

(meaning complete cure) against vivax malaria, and it remains today as the only such therapy with that licensed therapeutic indication.

As the American war in Vietnam loomed in the early 1960s, further clinical work on chloroquine and primaquine provided products useful to that military effort. The Americans developed methods of treatment and prophylaxis using these drugs that would not threaten the safety of its many G6PD deficient troops (the abnormality did not excuse them from eligibility for conscription and the US Army chose to not screen against it). They discovered that the most common African variant of G6PD deficiency (A-) could safely tolerate weekly doses of the drug over 8 weeks, as opposed to the relatively threatening daily dosing of 14 days. Remarkably, the efficacy of primaquine against relapse appeared to rest wholly upon total dose with almost no regard to schedule of that dosing; the same total dose of primaquine against *Plasmodium cynomolgi* in rhesus macaques – a superb model of vivax relapse – showed equally good efficacy when administered as a single dose, or multiple doses over weeks [98]. US military medicine thus delivered the “C-P” pill for protection of their troops against both acute attacks (with a weekly dose of 300 mg chloroquine) and relapse (with a weekly dose of 45 mg primaquine in the same tablet). As an added therapeutic bonus, it was shown that single 45 mg doses of primaquine in the company of chloroquine effectively killed the infectivity of the sexual stages responsible for malaria transmission [99, 100].

Thus, in the mid-1960s, corresponding with the nadir of global malaria in the wake of the GMEC, Northern S&T effectively abandoned malaria chemotherapy as a mission accomplished. That mission may be appreciated as having been driven by the threat of malaria to troops engaged in warfare in the tropics. The technically flawed and partial explorations that had been conducted nonetheless provided products suited to those Northern requirements and further or better explorations seemed pointless. The South would live with those chemotherapeutic solutions for the coming decades, and still does.

Chemotherapeutics Limbo: 1965–1995

The malaria chemotherapeutics status quo in 1965 and the reluctance of S&T in the North over the ensuing

30 years to effectively engage the problem leaves us where we stand today – poorly equipped to deal with the malaria problem of the South in a physical and strategic sense. An examination of the management of the highest chemotherapeutic priority – treating the acute attack – during this era, and our response to it, provides useful historic insights.

As has been described, the Northern R&D engine delivered chloroquine and primaquine around 1950 as the primary workhorse drugs for global malaria. Resistance to chloroquine by falciparum malaria emerged in the late 1950s and early 1960s. By the 1980s the problem had consolidated globally and new therapies trickled out of the few Northern laboratories engaged in chemotherapeutic discovery and development. Alternatively, older and less ideal drugs developed during the effort of the 1940s and 1950s were dusted off and put to work, for example, sufladoxine-pyrimethamine, proguanil, and others. However, resistance to those therapies emerged with astonishing speed and thoroughness [101]. Quinine combined with antibiotics like tetracycline or doxycycline then, in that crisis setting, found application against chloroquine-resistant falciparum malaria. Oral quinine, however, may be one of the most impractical therapies: it requires three doses daily for 7 days and causes a wide range of very unpleasant side effects; a syndrome called cinchonism that often deranges compliance to therapy.

Very few genuinely new therapies emerged during that era, most of them emanating from the US Army program at the Walter Reed Army Institute of Research (WRAIR) [102]. That program developed and fielded mefloquine (Lariam™) by the mid-1980s and it found broad application for chemoprophylaxis. However, patent protection, perceptions of its poor tolerability in many people, along with the emergence of resistance to it in areas of Southeast Asia, limited its broader application in almost all endemic zones. More recently, however, it has been partnered with an artemisinin, especially in the Mekong region, but not from GMP manufacturers. This exposes the drug to risk of poor manufacture and counterfeiting, and the packaging as separate pills risks poor compliance and onset of resistance to both artemisinin and mefloquine components. One co-formulated artesunate-mefloquine therapy has been registered in Brazil and other Latin American

countries through the Drugs for Neglected Diseases Initiative. Nonetheless, mefloquine may be considered a minor contribution to the toolbox against global malaria.

The other genuinely new therapy to emerge at the tail end of the same era (and with roots in WRAIR), atovaquone-proguanil (Malarone™) followed a similar path to global malaria irrelevance. Glaxo-Wellcome (later to become GlaxoSmithKline (GSK)) took up the cause of developing this drug for the prevention and treatment of malaria during the early 1990s. Their rapid success in registering it for these indications reflects both corporate determination and skill, and superior chemotherapeutic properties: very safe and well tolerated, highly effective against acute attacks of all the malarias (including multidrug resistant falciparum malaria), co-formulated, and available as a GMP product. The single known serious problem with this drug at the time, its relatively high cost, seemed to exclude it from the global malaria toolbox. GSK, however, generously aimed to see this extraordinary therapy put to work. In cooperation with various government and nongovernment partners, they created the Malarone Donation Program with the expressed aim of providing a million treatments with this drug annually at no cost to the end users [103]. However, flaws in this strategy would undo that program.

Evidence emerged that a single point mutation in *P. falciparum* caused it to be completely and irreversibly resistant to atovaquone + proguanil [104]. Malariologists worried that its broad application would quickly result in its loss, and perhaps other important drugs via cross-resistance [105]. The donation program coped with this potential danger by setting down very strict criteria for its clinical use. Consequently, at pilot sites in Kenya, only 0.68% of malaria patients received the treatment and that minority were among the least likely to suffer a fatal outcome [106]. Hence, mortality rates among the vastly excluded majority were not affected by the intervention. African authors of that “lessons learned” report [106] explained that new drugs were not the imperative; poor access to any drug drove transmission and high mortality. The complex social and political process that drove the donation program to pilot implementation in Africa excluded African voices wholly at first and later invited them but suppressed

their dissenting views [107]. The Northern voice and conviction that this new tool required implementation in the South, along with their fiscal authority and influence, dominated the strategic consensus. Despite valiant and conspicuously generous efforts by GSK to contribute to the global malaria problem with this drug, the drug plays a very minor role in malaria treatment and control in endemic zones.

Northern S&T between 1965 and 1995 effectively failed to yield a solution for the highest chemotherapeutic imperative in endemic zones: treatment of the acute attack. At Southern clinics around the globe in the 1990s and well into the 2000s patients were still being treated with drugs invented half a century earlier. These drugs were less safe and effective than chloroquine in the heyday of its efficacy, and a cost in human lives was rapidly accumulating [108]. This is when the contemporary efforts to field ACTs as front-line therapies, already described, at last took root and continue today. Treating the acute attack, one of several chemotherapeutic challenges with malaria, may be far less an S&T challenge than finding it, diagnosing it, and reporting it. Assuming that the diagnostics S&T challenge is met, have the chemotherapeutic tools that are needed to capitalize on the accurate identification of the infected person been assembled? Northern R&D obsession with the acute attack unfortunately comes with neglect of the other possibly useful drugs for managing endemic malaria at that critical time/place interface when the parasites meet the range of chemotherapeutic instruments aimed at them.

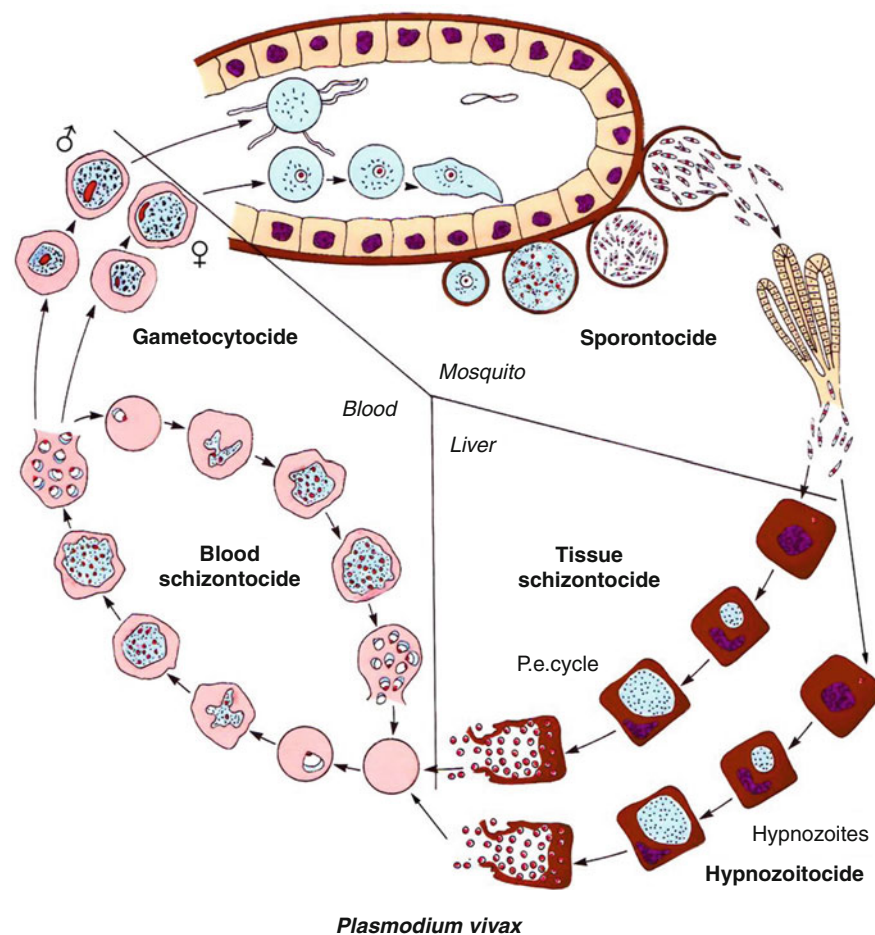
Chemotherapeutics Neglect

The chemotherapy of malaria consists of five distinct compartments, treatment of acute attack (using blood schizontocidal agents) being one. The other compartments include: drugs aimed against the dormant stages in liver, called hypnozoites, responsible for relapse (using hypnozoitocides); drugs aimed against active stages in the liver, called tissue schizonts, that lead to acute attacks (using tissue schizontocides); drugs affecting the sexual stages in blood, called gametocytes, responsible for infection of biting mosquitoes and transmission of malaria (using gametocytocides); and drugs that prevent the development of parasites in the mosquito gut (called sporontocides). [Figure 3](#) illustrates these compartments.

The blood schizontocidal antimalarials constitute the vast majority of antimalarial drugs in use today. These drugs, with few exceptions and when taken as directed, exert therapeutic activity only in that compartment. A few exhibit minor or limited activity elsewhere, for example, atovaquone may act as a tissue schizontocide against falciparum but not vivax malaria; chloroquine kills the gametocytes of vivax but not falciparum malaria; and the artemisinins will kill gametocytes of falciparum malaria but not if already mature. One drug in clinical use today remarkably exhibits therapeutic activity in all of these compartments. That drug is primaquine, and despite its extraordinary dexterity in killing a wide range of morphologically and physiologically distinct forms of parasites (including across species), it is a tool of very limited utility in the global malaria toolbox. This is a consequence of both failing to tame its sometimes dangerous side effects and to develop practical means of gauging its therapeutic efficacies.

G6PD Deficiency The hemolytic toxicity of primaquine, and the absence of a Northern R&D response to the problem, largely explains its very limited utility in endemic zones. Sustained daily dosing with primaquine (15 or 30 mg daily for 14 days) poses a potentially lethal threat to patients with some variants of G6PD deficiency [109]. The US Army developers of primaquine, however, effectively solved this problem by discovery of both the nonlethal threat of daily dosing in African-American soldiers with the A- variant of the deficiency, and by fielding the perfectly safe (in A- variants) and efficacious single weekly dose of 45 mg for 8 weeks [110]. The diagnosis of G6PD deficiency requires a cold chain for shipping and storage of reagents, specialized skills and equipment: micropipettors, water bath, test tubes, and ultraviolet lamp, for example. This poses little technical challenge to care providers in the North, but in the endemic zones of the South such laboratory capacities are very rare. The providers who most often see malaria must chance potentially serious harm or choose not to prescribe the drug. G6PD deficiency and the unsustainable nature of that diagnosis in endemic zones (as it has been for 50 years) render primaquine almost wholly ineffective as a therapy to prevent relapse.

Sites of Action of Antimalarial Drugs



Tropical Health and Sustainability. Figure 3

Life cycle of *P. vivax* and chemotherapeutic compartments. The life cycle of *P. falciparum* would be essentially similar to this with the lone exception of the hypnozoite shunt in the cycle, which that parasite lacks (Taken from reference [53] with the permission of the American Society of Microbiology. Figure by Prof. Wallace Peters with artwork by Ms. Andrea Darlow)

Development of a point-of-care diagnostic device for G6PD deficiency would substantially mitigate this problem. The US Army recently supported the development of such a tool in order to improve the safety and effectiveness of an experimental replacement for primaquine, another hemolytic 8-aminoquinoline called Tafenoquine [111]. The product of that effort, however, can only be used reliably at temperatures below 25°C [112]. Another commercial company in the USA (AccessBio©, New Jersey) independently took up the point-of-care challenge and has a product in research and development that is simple,

inexpensive, and heat-stable. That kit has performed well in initial field evaluations in Cambodia [113]. AccessBio© initiated this R&D effort with no support or guidance from mainstream malaria R&D funders, and WHO provided the funding for that evaluation (JKB is an investigator on that effort).

Primaquine suited Northern settings and requirements, and adapting it to safe use in Southern settings has never been an R&D imperative. Among the many dozens of G6PD deficiency variants, the primaquine sensitivity phenotype has been characterized in only three geographic settings (African A-, Mediterranean

B-, and, only recently, Mahidol from Thailand). These variants represent mild, severe, and moderate primaquine sensitivity phenotypes, respectively. Although residual G6PD enzyme activity has been characterized in most variants, its use as a surrogate for the primaquine sensitivity phenotype is a supposition not supported with clinical evidence. WHO has long recommended a single 45 mg dose without screening for G6PD deficiency, but largely upon the (inadequate) safety evidence from the USA that involved healthy African-American volunteers having the mild A- variant of this disorder [94]. However, in a very limited series of studies in the USA and Italy, a single 45 mg dose caused acute intravascular hemolysis of 25% of red blood cells in otherwise healthy volunteers having the Mediterranean B- variant (severe primaquine sensitivity) [114, 115]. Safe dosing in one variant may not hold for others.

Gametocytocide Poor understanding of primaquine sensitivity phenotypes creates very serious practical problems in malaria control. This may be most appreciated by the application of a single 45 mg dose of primaquine as a gametocytocide against falciparum malaria [116]. A 25% hemolysis in an acutely ill and already anemic patient may contribute significantly to onset of life-threatening complications. Remarkably, the safety of primaquine gametocytocidal therapy in patients with malaria, much less G6PD deficiency variants of any type, has not been systematically carried out. Limited recent studies from Africa suggest a significant threat to malaria patient safety [117, 118]. Countries aiming to apply the 45 mg dose of primaquine as a supplement to their control or elimination strategies must first demonstrate safety among the variants of G6PD deficiency occurring in their areas of operations. In 2011, the WHO undertook sponsorship of such studies in Cambodia using its own relatively meager R&D funding resources (JKB is an investigator on that effort).

The safety and efficacy of primaquine as a gametocytocide in falciparum malaria were evaluated with chloroquine therapy for the acute attack over 50 years ago [99, 100]. Chloroquine is no longer used in this capacity, and no clinical evaluation of the safety and efficacy of primaquine as a gametocytocide when administered with any ACT has occurred.

The artemisinins exert activity against immature gametocytes and could possibly yield a lower and less threatening optimal dose of primaquine. Northern R&D, however, expresses no agenda for clinical research on primaquine for any therapeutic indication.

The most conspicuous neglect of the gametocytocidal compartment has been the failure to conduct a systematic chemical survey of compounds for this activity. Northern S&T obsession with the acute attack and neglect of other avenues to address the problem of endemic malaria may be most evident in this line of investigation. An examination of the fairly extensive body of work on gametocytocidal properties of chemotherapeutic agents reveals that almost all such evaluations have been conducted on drugs with proven therapeutic activity against the asexual blood stages causing acute malaria [119]. Killing gametocytes has apparently been thought of strictly as a therapeutic bonus to treatment of the acute attack. The possibility of a safe, effective, and perhaps already widely available drug at low cost providing a means to effectively arrest transmission has not been explored. Northern S&T could have, with relative technical ease, conducted such explorations over 35 years ago. A 2011 call for proposals for such work from BMGF represents a welcome sign of dawning realization of the importance and enormous promise of this approach.

Hypnozoitocide The primary therapeutic application of primaquine, as a hypnozoitocide, offers further evidence of Northern R&D inability to respond effectively to Southern problems with critical tools in its chemotherapeutics toolbox. Although primaquine has been the only therapeutic agent against relapse for the past 60 years, almost no research of its mechanisms of activity against the hypnozoites has been carried out. There is no understanding of its therapeutic target, much less which of its many metabolites may be most active. Resistance to primaquine by hypnozoites almost certainly occurs, but there is no validated means of confirming that phenotype and complete ignorance of the corresponding genotypes. Simply observing responses to therapy does not adequately address that problem because parasitemia occurring weeks to months after treatment may represent either therapeutic failure (relapse despite therapy) or simply a new infection. No means of distinguishing those events by

molecular genotyping has been established, largely due to the biological complexity imposed by the hypnozoite [120]. Unambiguous estimates of primaquine efficacy thus require the absence of risk of reinfection. Such studies have not been done since the development of primaquine in experimentally challenged American prisoners in the 1940s and 1950s [121]. Consequently, there is almost no grasp of the therapeutic efficacy of the only drug against relapse in vivax malaria anywhere it occurs. Primaquine may already have been lost to resistance and there is no validated means of evaluating that possibility.

Chemotherapeutics Performance Evaluation

Over the past six decades, through periods of ebbing and flowing fiscal support, Northern R&D has focused available energies and resources, and still does, on the acute attack of falciparum malaria. The almost complete neglect all other forms of chemotherapy currently in the toolbox of the Southern craftsman working the problem now imperils successful elimination goals. And we were warned as L.T. Coggeshall [122] wrote in 1952, “*The greatest misconceptions about the treatment of malaria, especially in the past, has arisen from the fact that too many considered it a single disease. Malaria is not a disease – it is a variety of diseases.*” There are therapies for one of the malaras and, effectively, almost none for the others. One key element of this neglect has been the failure to adapt *P. vivax* to continuous in vitro cultivation in Northern laboratories as occurred with *P. falciparum* in the 1970s [123]. Setting aside the improbability of adequate R&D investment in achieving that feat for *P. vivax*, this failure does not excuse Northern R&D from dealing with this parasite. While it is true the problem limits Northern laboratories from effectively engaging the problem in the North, transfer of the required R&D capacities southward stands as an obvious but inadequately funded solution. In this and other issues of chemotherapeutics neglect outlined here, it remains unclear whether the 3.1% share of malaria R&D spending on *P. vivax* [8], and perhaps even less on gametocytocidal therapies, is the cause or effect of their neglect. Either way, “inadequate” effectively describes Northern R&D performance in these vital pieces of the malaria control toolbox.

In a chemotherapeutics sense, the Southern craftsman trying to disassemble endemic malaria may be thought of as equipped with a fixed-gauge wrench for nuts of many gauges.

The Vaccine Development Gambit

In 2011, a phase III clinical trial of a malaria vaccine called RTS, S in 15,640 children aged up to 17 months is underway at 11 study sites in Africa [124]. This bold and expensive effort reflects the enormous technical and fiscal challenges in malaria vaccine development. It also expresses the Northern determination to produce this particular tool despite the extraordinary gamble of resources and will.

The leading edge biotechnology-driven effort to develop vaccines against malaria has been an epic, contentious, and, thus far anticlimactic journey of 30 years [125]. The modern effort to engineer a protein subunit vaccine began in earnest and with a great deal of optimism in the early 1980s with emergent recombinant DNA technology swinging open a very promising door. Having genetically engineered bacteria to produce malaria proteins in kilogram quantities seemed a means of conquering malaria with relative ease. Instead, through a long series of preclinical and clinical R&D efforts around the globe, *P. falciparum* has consistently shrugged off the eloquently derived antibodies and T-cells arrayed against it [125–127].

The steepness of the R&D challenge in vaccination stems from the failure to adequately understand the enormously complex immune response provoked by malaria. That response varies terrifically across variables like age, history of exposure, intensity of exposure, and host and parasite genetics [128]. The dawn of the scientific understanding of immunity to malaria began in 1899 when Robert Koch visited Dutch colleagues at what would become the Eijkman Institute in Batavia on the island of Java in the East Indies. Koch noticed that densities and frequencies of parasitemia declined sharply with age at heavily endemic Ambarawa in central Java, whereas at much less endemic Sukabumi in western Java those values remained relatively level across ages. Koch surmised that heavy exposure to infection, over a period of many years, imbued protection against high levels of parasitemia and attendant disease (see citations

in [128]). The same pattern has been consistently observed over the decades and naturally acquired immunity has been considered the cumulative product of chronic and heavy exposure to infection. Immunologists envision the human immune system “learning” the antigenic repertoire of the parasite by repeated exposures and eventually, at around adolescence, mastering it [129].

Likewise, vaccinologists viewed their challenge as one of “teaching” that immune response by exposure to the proper parasite antigen or set of antigens. Though very simply expressed, this characterizes the malaria vaccine development strategy. Accepting the supposition that very young children in heavily endemic Africa constitute the majority of deaths due to malaria, the expressed R&D priority of vaccinating them is sensible. This demographic group is indeed vulnerable, exceedingly so compared to their parents. Adult Africans in heavily endemic areas, though almost constantly infected, very rarely suffer even minor illness (unless pregnant, especially for the first time). The susceptibility of their children to anemia, cerebral malaria, pulmonary complications, and other life-threatening syndromes stands in stark contrast [130]. They seem to have not “learned” to tame this infection and the appropriately “instructive” vaccine would mitigate their losses to malaria. This simplistic concept suffers logical pitfalls, however, evident with closer examination of the dynamics at work between those very young hosts and the parasites invading them.

A single episode of malaria in a nonimmune traveler, left untreated for too long, very often threatens rapid devolution into a life-threatening event. African children, on the other hand, experience about six to eight new infections each year in many endemic zones. Over 95% of them survive this rigorous challenge and go on to become malaria-hardy adults. This survival, surely impacted by access to therapy, nutrition, genetics, and probably many other factors, must also reflect successful immunological management of the always-present threat of sliding into lethal complications [128]. The loss of some of these children to malaria begins to less resemble a failure to “learn” a specific immune response from its parasite instructor, but more like an intrinsic failure to mount the appropriate immune response that permitted survival among almost all others in their cohort. And yet, vaccine R&D

remains focused on parasite antigens rather than the intrinsic basis for success versus failure in what amounts to a 95% efficacious vaccination by the parasite itself. There is almost no understanding of what constitutes an appropriate versus inappropriate immune response to natural infection and the determinants of that life versus death turning of events. This gap in understanding extends to responses to experimental vaccines.

The consequences of such poor understanding may be found embedded in the process of malaria vaccine development and it imposes extraordinary financial risk. Unlike drug development, where compounds may be rejected or pushed forward on the basis of unambiguous laboratory determinations of activity against the parasite, malaria vaccine developers possess no such tool of discernment. No single immune response or set of responses has been shown to reliably predict protection in the infected human at any age or history of exposure, much less specifically infants and small children chronically exposed to infection. Vaccine candidates advance (or do not) through development algorithms on the basis of rational suppositions about protection in humans. Immunogenicity in terms of specific humoral or cellular targets, for example, may be the basis of development decisions even though those responses may have no evidence-validated role or compelling and consistent correlation with achieved protection in humans. Phase I and II trials in healthy nonimmune European or American adults for vaccines aimed at African infants and effectively already immune children compounds this problem. These generally acknowledged pitfalls [8] effectively defer credible vaccine development decisions to phase II and III trials. The vaccine RTS,S, now in phase III trials, was invented in 1987, tested in humans in 1992, evaluated in Africans in 1995, and evaluated in small numbers of African children between 2004 and 2007. The inadequate understanding of immunity to malaria as it occurs in the real world imposes a vaccine development paradigm resembling a roulette wheel with each turn of it coming with decades of work and many millions of dollars. Surely only repeated success could sustain such a development paradigm, and yet it is one having 30 years of failure. What sustains it?

R&D Investment

Today, malaria vaccine proponents glare over their fiscal parapets at their antimalarial drug development competitors vying for the same large share of malaria R&D resources. Each points to the failings and shortcomings of the other in making their cases for funding before donors and sponsors. The vaccines camp usually, albeit narrowly, wins the resources contest (Fig. 1). The reasons for the consistency of this large investment, despite decades of frustration, perhaps reflect how we perceive the malaria problem and, arguably, national self-interest in S&T capacity, agility, and economic spin-off.

Malaria vaccine development has and continues to probe the very far leading edges of contemporary and potentially industrial biotechnological immunology [131]. Beginning in earnest in the 1980s, the best and brightest were generously resourced and engaged in a globally important scientific endeavor. Northern laboratories blossomed and hummed with activity [127]. The US and European biotechnology industries enjoyed many technological and economic spin-offs from the effort. One of the malaria vaccine biotechnology pioneers, Wayne T. Hockmeyer, retired from the US Army in the late 1980s and went directly on to found Molecular Vaccines Inc. (later renamed MedImmune), a biotechnology concern purchased by AstraZeneca in 2007 for \$15.2 billion. The vaccines industry saw and participated in proofs-of-concept of novel approaches to vaccine development like DNA vaccines and arrays of sophisticated adjuvants. In short, over 30 years of malaria vaccine development produced many scientific, technological, and economic dividends without delivering, yet, a licensed vaccine.

This is the nature of science, where exploration leads to unexpected destinations. Simply having a technical problem and the wherewithal to address it is key to S&T success even without directly solving the primary problem. The nature of funding in malaria vaccine development illustrates the linkage between S&T capacity and national S&T self-interest. The laboratories that executed malaria vaccine development work were supported by the nations where the laboratories were located, with few minor exceptions. American-funded efforts went primarily to American laboratories, as was true of European efforts, and even

one bold effort in Colombia [132] was funded by that government. What is likely at work is the irresistible drive to reap the peripheral benefits of S&T endeavors within borders. No nation in the endemic zones, save the valiant Colombian example, undertook vaccine discovery and development. That may be due to a lack of fiscal resources, standing technical capacity, and an available technologically endowed workforce. But those determinants themselves provide grim reflection of the unwillingness of Northern nations, collectively, to make substantial direct investment in S&T foundations in the South.

It may be argued that the North harvests S&T/R&D dividends derived from a Southern problem – and reasonably counter-argued that the investments are their own and, because of standing capacities, the less costly and speediest avenue to solving the problem. The Northern R&D posture would be supportable if it delivered the most useful tools, but it has not and does not seem strategically positioned to do so. Northern R&D engaging the malaria problem yields tremendous and important advances in S&T, but it historically exhibits very limited abilities in delivery of practical and useful tools that significantly impact malaria as a global health problem.

Sustainable S&T in Tropical Health

Northern R&D remains the technological fulcrum for the public health lever against Southern infections, and this may be an important contributor to stymied and perhaps unsustainable progress. Gordon Harrison wrote in his 1978 history of malaria [64] of the failure of the global eradication effort in the prior decade, “*Failure so universal, so apparently ineluctable, must be trying to tell us something. The lesson could be of course that we have proved incompetent warriors. It could also be that we have misconstrued the problem.*” The failure of Northern R&D agendas to field relevant tools against endemic Southern malaria must also be trying to tell us something. Bench malaria science of Northern S&T is a powerful engine of scientific discovery and human progress engaged in the fight against malaria, and the aim here is not to demean those earnest endeavors by compassionate and dedicated people. The aim is to spark reconsideration of reliance solely upon that engine for practical solutions.

Southern scientists may be criticized for having abandoned the field to Northern scientists, but this is unjust and self-serving of the critic. As the Malarone Donation Program already described illustrates, Northern bias, imperatives, and fiscal influence may displace Southern voice and initiative out of the strategic equation. This is where the Northern perception of a lack of Southern awareness of useful technologies plays in the unhelpful assumption of Northern leadership on decisions regarding tools intended for Southern use. In the Malarone example, Northern science perceived a powerful new drug available at no cost, whereas African stakeholders perceived a futile intervention not addressing the real problem. Southern scientists put up futile active and passive resistance to the program because the proposed intervention deflected alternatives and the prospect of progress. The report of Shretta et al. [110] documents this history and should be studied by Northern scientists engaging the malaria problem.

What the Southern scientists pointed to was not the drug, but the availability of diagnosis and treatment where most deadly malaria occurred [109, 110]. It may be that Northern S&T perceives such a problem as one it is not particularly suited to engage. In other words, the problem seems practical rather than technological, at least in how the North perceives its technology imperatives. The Northern perception may be that problems that do not engage the suite of Northern technology assets need not involve them. That may be, but their technological imperatives and financial resources tend to suffocate alternative approaches by Southern scientists co-opted into a Northern R&D agenda.

Malaria vaccines represent the primary Northern S&T answer, or at least hope, to solve lethal malaria at isolated sites all across the global endemic belt. Vaccination in lieu of very hard to sustain diagnosis and treatment services seems the ideal solution. Vaccines have pushed the burden of morbidity and mortality due to infections to the lowest point in all of human history, and such a track record is hard to argue against. However, one may forcefully argue that hope is a poor offensive strategy against entrenched malaria. Determination, will, resources, and faith in the success of vaccines may be defied by the unfolding of an unknowable future. Preparing for

possible failure of vaccines neither endorses nor ensures it, and responsibly envisioning a future without vaccines prudently commands attention to alternative technological approaches. Putting the development of those approaches into Southern hands may be the surest means of success with most appropriate technologies providing sustainable solutions.

Southern Scientific and Technological Development

This entry has emphasized the problems with Southern reliance upon Northern technological innovation to diminish preventable endemic infectious disease burdens. A South empowered with indigenous S&T capacities would likely generate more effective instruments that more quickly lift those burdens. Such capacity would likewise, as it did in the North, spin off substantial economic benefits that indirectly drive endemic diseases down and health up. Achieving this would require deliberate action by those nations of the North capable of delivering such capacities, and the foresight to see very significant short-term losses recovered many times over in the longer term.

It may be argued, despite the MIM example already discussed, that the North already invests in the development of S&T capacities of the South. Each year many Southern scientists receive academic and technological training in the North with generous support from Northern sponsors in the government, academia, and private sectors. This approach, however, may be unlikely to yield real capacities in Southern S&T. Many of these students acquire skill sets that cannot be applied in their home nations. There are few institutions with the technological capacities of the Northern universities they attend. These students typically either remain in the North as valued technologically skilled immigrants, or they go home and assume positions that do not hone and deepen their technological abilities. Many of the latter subsequently relocate to developed nations selfishly promoting immigration policies designed specifically to lure them away [133]. The promise of advanced S&T training for Southern scientists in the North rings hollow without Southern R&D institutions where those skills may be put to effective work.

The key may be establishing S&T “nurseries” in the South having specific technological suites and focused objectives. In the example of malaria, few could argue against the potential direct and indirect benefits of an institution committed wholly to the development of medically useful insecticides or repellants, or another dedicated to a thorough chemical exploration of potential therapeutic agents aimed solely at sexual blood stages and arresting malaria transmission. The production of GMP malaria vaccines would nurture a commercial industrial biotechnological base capable of expansion to other vaccine-preventable illnesses and pharmaceuticals. S&T sustainability successes standing today in the South offer examples.

The Eijkman Institute for Molecular Biology in Jakarta, Indonesia is named after the Dutch scientist who discovered vitamin B as the cause and cure of beriberi with work executed largely in that city between 1886 and 1925. Christiaan Eijkman was awarded the Nobel Prize in 1929 for the discovery of vitamins [134]. During this heyday, the institute employed both Dutch and Indonesian scientists functioning as full peers heading their own avenues of biomedical research of tropical infections. The war in the Pacific saw the Dutch scientists arrested and imprisoned. The Indonesian scientists carried on with the work of the institute largely without direct interference from the Japanese forces. Tragically, however, an occupier suspicious of intellectual elites executed many Indonesian doctors and medical scientists [135]. The director of the Eijkman Institute in 1944, Achmad Mochtar, offered a false confession of the murder of 900 Indonesian slave laborers by sabotage of a vaccine administered to them. He did this in exchange for the liberty of almost the entire Indonesian technical staff of the Eijkman Institute being held in prison and brutally tortured. In July 1945, the occupiers executed Mochtar. This event, and the bloody war for independence from the Netherlands that immediately followed, effectively destroyed the institute and it formally closed in 1965 [136].

In 1990, the dynamic then Minister of Science and Technology (later to become the third President of Indonesia), B.J. Habibie, devoted considerable government resources and his formidable force of personality to reopen the institute as a nursery for biotechnological innovation. He segregated it from the ministerial hierarchies in order to protect it from the political vagaries

of the Indonesian federal system. The laboratory became a semiautonomous institutional appendage to his ministry. Habibie, himself a product of the European Airbus Industrie S&T engine, sought an Indonesian scientist immersed, tried, and tested in the culture of Northern S&T. He found and recruited Prof. Sangkot Marzuki, then a tenured and highly productive cellular and molecular biologist at Monash University in Australia. The fledgling laboratory in Jakarta received generous support from the government until the economic collapse and political turmoil of 1998. The laboratory nonetheless survived by its own scientific and technological wits, becoming a largely self-sustaining research institution through the conduct of work paid for by competitive, merit-driven grants. Their focus on molecular biology endows them today with sophisticated gene sequencing and analysis capacities. Those are being leveraged to attract funding for specific work on discrete pathogens endemic to Indonesia, in addition to a variety of human genetic, wildlife conservation, and even criminal law enforcement molecular biological problems.

The primary point with this example is that young Indonesian scientists, whether trained at home or abroad, have a stable institution where they may sharpen their skills with ever-steeper technological challenges under the mentorship of both senior Indonesian scientists and their foreign partners. The scientists in the institute are not government employees and their salaries are thus neither guaranteed nor constrained. They are able to commit fully to the S&T mission, with the certain understanding that secure employment demands scientific productivity and overcoming stiff competition from other labs engaged in similar work around the world. More importantly, consistent success provides a career of advancement in abilities, responsibilities, and compensation reflective of such success.

Almost all scientists today manage their professional lives in this way, no matter where they happen to live. Doing science at the leading edge is a hard-earned privilege reserved for those able to demonstrate the ability to contribute to the furthering of that edge. Instilling this understanding, and providing places where sustained success is possible and deliberately nurtured, substantially furthers the S&T capacity of the Republic of Indonesia as a whole. It is not the

S&T per se that makes this possible; it is the institute, its foundations, and its culture of uncompromising scientific excellence. B.J. Habibie and the leadership of the Eijkman Institute understood this and achieved it.

That institute is certainly not the only example of significant progress in the deliberate long-term commitment to the systematic nurturing of Southern S&T capacities. The Pasteur Institutes around the world provide excellent nurseries for local S&T, as do the laboratories funded by the Wellcome Trust in Kenya, Tanzania, Thailand, and Vietnam. Indeed, the former Pasteur Institute (founded in 1896) at Bandung, West Java, created most of the vaccines used in Southeast Asia prior to the Second World War, and the foundations of that investment remain today in the same physical location occupied by the government-owned industry that produces most of Indonesia's vaccines. Not coincidentally, Indonesia's vibrant and lucrative pharmaceutical manufacturing industries surround that facility. Bringing S&T capacities southward obviously has and can provide sustainable S&T and R&D enterprises and their biotechnological industrial spin-offs.

The North must overcome its impatience and evidently misplaced self-confidence in its ability to effectively engage Southern health problems. S&T is the key to successfully resolving those burdens, but probably not as practiced in the North. Supplanting Southern S&T into the leadership role of R&D aimed at its infectious disease burdens should be deliberately planned, financed, and executed.

Future Directions

The certain failures and likely strategic missteps by Northern S&T put forth in this entry aim at dispassionate and rational reflection on the means of addressing Southern health problems in the S&T/R&D arena. Accepting Northern R&D agendas as poorly suited to the task, despite the humanitarian urgency, brings appropriate and useful focus to mitigating deficiencies in Southern R&D capacities. Resisting the fiscal and temporal expediencies that push R&D capacities and agendas northward may prove the surest means of sustainable and substantial gains in tropical health. Placing durable R&D capacities in the South as a primary rather than tangential benefit of fruitful North-South S&T

collaboration generally characterizes a direction aiming at that envisioned future.

Southern scientists have much to learn from their Northern counterparts, especially those in positions of institutional leadership of R&D endeavors. The intensely competitive leading edge of S&T requires institutional foundations that create and reward healthy scientific competitiveness. Northern R&D laboratories that fail to compete quickly vanish, and the cruel Darwinian-like forces that shape effective Northern R&D must also be brought to bear on Southern R&D enterprises – their vitality and sustainability requires it. Transfer of those R&D survival skills is perhaps where Northern S&T may be put to most effective use in the South. And the task of transfer should not be viewed as passive, as by example, but direct and deliberate participation in the building of those skills and capacities in Southern R&D enterprises. Acknowledging limited Southern S&T capacities as the core impediment to improving Southern health, rather than possible or hoped for tools emanating from Northern S&T, naturally leads in that direction.

Such a course requires strategic thinking, long-term commitment, flexibility, and inspired international cooperation and trust from both North and South partners undertaking it. Northern investment ought to aim at institutional R&D competitiveness rather than the R&D agenda itself. Logically, however, this inevitably leads to establishing successful R&D endeavors of a specific goal-oriented nature. Scientific productivity would be engrained in any rational North-South institutional R&D capacity-building cooperative endeavor. In short, success in the core institutional capacity-building objective necessarily carries with it R&D success and rewards for both partners. An effective win-win formula ensures their mutual R&D survival, and therefore commitment to the endeavor. Perhaps most importantly, the Northern partner would not be bringing to the partnership a predetermined R&D agenda. That would be decided between the partners and the sponsor of their endeavor.

That third party in this envisioned collaboration, the sponsor, represents the primary challenge in realizing it. At least in the realm of malaria R&D today, that sponsor is most likely to be an American funding agency with what it views as clear product- or knowledge-development paths and technical milestones to accomplish in achieving success, that is,

fielding a product that it decided would be useful. The Northern obsession with the development of products of its conception for Southern health progress should be acknowledged as a significant impediment and tempered if not abandoned.

The assumption of Southern R&D leadership on Southern health issues would require many years of determined will and real reform on both sides of the Tropic of Cancer to achieve and sustain. It may require pragmatic lessening of the faith Northern R&D exhibits in its ability to deliver solutions in the short term. Southern awareness of the intensely competitive nature of biomedical R&D should rise to a level consistent with firm commitment to that very difficult course. Southern R&D enterprises must not depend wholly upon funding from noncompetitive sources, and gauge their own relevance upon the objective evidence of competitive survival.

Southern public and private donors also need to participate and engage R&D endeavors managed by Southern institutions. Northern dominance of the R&D agenda in part stems from the reluctance of Southern donors to make these investments and thus bring that voice and influence to discussion and consensus. Their absence becomes ever more conspicuous with the booming economies of East, South East, and South Asia in comparison to the now troubled Northern economies. Establishing stable and sustainable R&D enterprises in the South may imbue such donors with the confidence and even self-interest to make such investments. Northern investments in Southern R&D should at least be matched by Southern governments and donors positioned to contribute. The South desperately requires humanitarian visionaries and champions, as in the models of Sir Henry Wellcome, Bill and Melinda Gates, Warren Buffett, and others in the North, to generously and intelligently unleash available capital upon very serious tropical health challenges.

North-South S&T/R&D cooperation on tropical health, after all, aims at lifting the mingled burdens of endemic diseases and poverty in the South. Northeast Asia largely achieved this feat over the past 50 years and today is the mercantile engine of the global economy. Brazil in South America has made large strides in health and trade and with it drives a substantial portion of the North American economy. Healthy people liberated

from the pursuit of survival assume the pursuit of happiness that drives modern economies. They become educated, economically creative, and productive, selling to each other the accoutrements of easier and more fulfilling lives. Acknowledging North-South S&T/R&D parity as a key element in such progress moves far broader development agendas rapidly forward. Success in these endeavors may ultimately remove the conceptual utility of dividing the world into these two disparate spheres.

Acknowledgments

JKB is supported by the Wellcome Trust grant #B9RJIXO. Prof. Jeremy Farrar, Dr. Simon Hay, and Dr. Trevor Jones provided helpful reviews of the manuscript.

Bibliography

Primary Literature

1. Galama T, Hosek J (2008) U.S. competitiveness in science and technology. RAND Corporation, Santa Monica
2. Committee on Ensuring the Best Presidential and Federal Advisory Committee Science and Technology Appointments, National Academy of Sciences, National Academy of Engineering, Institute of Medicine (2005) Science and technology in the national interest: ensuring the best presidential and federal advisory committee science and technology appointments. The National Academies Press, Washington, DC, p 208
3. Lu C, Schneider MT, Gubbins P, Leach-Kemon K, Jamison D, Murray CJL (2010) Public financing of health in developing countries: a cross-national systematic analysis. *Lancet* 375:1375–1387
4. Sachs J, Malaney P (2002) The economic and social burden of malaria. *Nature* 415:680–685
5. WHO (2010) World malaria report 2010. WHO, Geneva
6. Hay SI, Okiro EA, Gething PW et al (2010) Estimating the global clinical burden of *Plasmodium falciparum* malaria in 2007. *PLoS Med* 7:e1000290
7. Gething PW, Smith DL, Patil AP, Tatem AJ, Snow RW, Hay SI (2010) Climate change and the global malaria recession. *Nature* 465:342–345
8. PATH (2011) Staying the course? Malaria research and development in a time of economic uncertainty. PATH, Seattle
9. Hay SI, Guerra CA, Tatem AJ, Noor AM, Snow RW (2004) The global distribution and population at risk of malaria: past, present, and future. *Lancet Inf Dis* 4:327–336
10. NIH (2008) NIAID agenda for malaria, Bethesda, Maryland
11. WHO (2008) Global Malaria Action Plan, Geneva
12. Nosten F, Luxemburger C, ter Kuile FO, Woodrow C, Eh JP, Chonsuphajaisdhi T, White NJ (1994) Treatment of

- multi-drug resistant *Plasmodium falciparum* with a 3-day artesunate-mefloquine combination. *J Infect Dis* 170:971–977
13. WHO (2008) World malaria report. WHO, Geneva
 14. Snow RW, Guerra CA, Noor AM, Myint HY, Hay SI (2005) The global distribution of clinical episodes of *Plasmodium falciparum* malaria. *Nature* 434:214–217
 15. WHO (2009) World malaria report. WHO, Geneva
 16. Hay SI, Okira EA, Gething PW, Patil AP, Tatem AJ, Guerra CA, Snow RW (2010) Estimating the global clinical burden of *Plasmodium falciparum* malaria in 2007. *PLoS Med* 6:e1000290
 17. Elyazar I, Hay SI, Baird JK (2011) Malaria distribution, prevalence, drug resistance and control in Indonesia. *Adv Parasitol* 74:41–175
 18. Harijanto PN (2010) Malaria treatment using artemisinin in Indonesia. *Acta Med Indones* 42:51–56
 19. Dhingra N, Jha P, Sharma VP et al (2010) Adult and child malaria mortality in India. *Lancet* 376:1768–1774
 20. Kochar DK, Saxena V, Singh N et al (2005) *Plasmodium vivax* malaria. *Emerg Inf Dis* 11:132–134
 21. Kochar DK, Tanwar GS, Khatri PC et al (2010) Clinical features of children hospitalized with malaria - a study from Bikaner, Northwest India. *Am J Trop Med Hyg* 83:981–989
 22. Srivastava S, Ahmad S, Shirazi N, Verma SK, Puri P (2011) Retrospective analysis of vivax malaria patients presenting to tertiary care referral centre of Uttarakhand. *Acta Trop* 117:82–85
 23. Sharma A, Khanduri U (2009) How benign is benign tertian malaria? *J Vector Borne Dis* 46:141–144
 24. Barcus MJ, Basri H, Picarima H et al (2007) Demographic risk factors for severe and fatal vivax and falciparum malaria among hospital admissions in northeastern Indonesian Papua. *Am J Trop Med Hyg* 77:984–991
 25. Genton B, D'Acremont V, Rare L et al (2008) *Plasmodium vivax* and mixed infections are associated with severe malaria in children: a prospective cohort study from Papua New Guinea. *PLoS Med* 5:e127
 26. Tjitra E, Anstey NM, Sugiarto P et al (2008) Multidrug-resistant *Plasmodium vivax* associated with severe and fatal malaria: a prospective study in Papua, Indonesia. *PLoS Med* 5:e128
 27. Andrade BB, Reis-Filho A, Souza-Neto SM et al (2010) Severe *Plasmodium vivax* malaria exhibits marked inflammatory imbalance. *Malar J* 9:13
 28. Beg MA, Khan R, Baig SM et al (2002) Cerebral involvement in benign tertian malaria. *Am J Trop Med Hyg* 67:230–232
 29. Lomar AV, Vidal JE, Lomar FP et al (2005) Acute respiratory distress syndrome due to vivax malaria: case report and literature review. *Braz J Infect Dis* 9:425–430
 30. Lawn SD, Krishna S, Jarvis JN, Joet T, Macallan DC (2003) Case reports: pernicious complications of benign vivax malaria. *Trans R Soc Trop Med Hyg* 97:551–553
 31. Spudick JM, Garcia LS, Graham DM, Haake DA (2005) Diagnostic and therapeutic pitfalls associated with primaquine-tolerant *Plasmodium vivax*. *J Clin Microbiol* 43:978–981
 32. Sternberg GM (1884) Malaria and malarial diseases. Wood, New York, p 329
 33. Kitchen SF (1949) Vivax malaria. In: Boyd MF (ed) *Malariology*, vol II. W.B. Saunders, Philadelphia
 34. Ewing J (1902) Contribution to the pathological anatomy of malarial fever. *J Exp Med* 6:119–180
 35. Bassat Q, Alonso P (2011) Defying malaria: fathoming severe *Plasmodium vivax* disease. *Nat Med* 17:48–49
 36. Eldridge WW et al (1925) Treatment of paresis: results of inoculation with the organism of benign tertian malaria. *JAMA* 84:1097–1101
 37. Ferraro A et al (1927) The malaria treatment of general paresis. *J Nerv Ment Dis* 65:225–239
 38. Freeman W, Eldridge WW, Hall RW (1934) Malaria treatment of dementia paralytica: results in 205 cases after five to eleven years. *South Med J* 27:122–126
 39. O'Leary PA, Welsh AL (1933) Treatment of neurosyphilis with malaria: observations on nine hundred and eighty-four cases in the last nine years. *JAMA* 101:498–501
 40. Fong TCC (1937) A study of the mortality rate and complications following therapeutic malaria. *South Med J* 30:1084–1088
 41. Krauss W (1932) Analysis of reports of 8,354 cases of IMPF-malaria. *South Med J* 27:537–541
 42. Nicol WD (1932) A review of seven year's malarial therapy in general paralysis. *J Ment Sci* 78:843–866
 43. James SP, Nicol WD, Shute PG (1936) Clinical and parasitological observations on induced malaria. *Proc R Soc Med* 29:27–42
 44. Verhave JP (2010) personal communication
 45. James SP, Nicol WD, Shute PG (1932) A study of induced malignant tertian malaria. *Proc R Soc Med* 25:37–52
 46. Handayani S, Chiu DT, Tjitra E et al (2009) High deformability of *Plasmodium vivax*-infected red blood cells under microfluidic conditions. *J Infect Dis* 199:445–450
 47. O'Donnell J, Goldman JM, Wagner K et al (1998) Donor-derived *Plasmodium vivax* infection following volunteer unrelated bone marrow transplantation. *Bone Marrow Transplant* 21:313–314
 48. Finkel M (2007) Bedlam in the blood: malaria. *National Geographic*, July 2007, pp 32–67
 49. UN Millennium Project (2005) Coming to grips with malaria in the new millennium. Earthscan, London, p 129
 50. US CDC (2011) Grand rounds: the opportunity for and challenges to malaria eradication. *MMWR* 60:476–480
 51. Guerra CA, Howes RE, Patil AP et al (2010) The international limits and population at risk of *Plasmodium vivax* malaria in 2009. *PLoS Negl Trop Dis* 8:e774
 52. Baird JK (2004) Resistance to chloroquine by *Plasmodium vivax*. *Antimicrob Agents Chemother* 48:4075–4083
 53. Baird JK (2009) Resistance to therapies for *Plasmodium vivax*. *Clin Microbiol Rev* 22:508–534
 54. Baird JK (2005) Effectiveness of antimalarial drugs. *N Engl J Med* 352:1565–1577
 55. Baird JK (2011) Resistance to chloroquine unhinges vivax malaria therapeutics. *Antimicrob Agents Chemother* 55:1827–1830

56. WHO (2009) Rapid diagnostic test performance: results of WHO product testing of malaria RDTs: round 1 (2008). WHO, Geneva
57. Epstein J, Rao S, Williams F et al (2007) Safety and clinical outcome of experimental challenge of human volunteers with *Plasmodium falciparum*-infected mosquitoes: an update. *J Infect Dis* 196:145–154
58. Syafruddin D, Krisin, Asih P et al (2009) Seasonal prevalence of malaria in West Sumba district, Indonesia. *Malar J* 8:8
59. Harris I, Sharrock WW, Bain LM et al (2010) A large proportion of asymptomatic *Plasmodium* infections with low and sub-microscopic parasite densities in the low transmission setting of Temotu Province, Solomon Islands: challenges for malaria diagnostics in an elimination setting. *Malar J* 9:254
60. WHO (2011) Consideration of mass drug administration for the containment of artemisinin resistant malaria in the Greater Mekong Sub-Region. Report of a consensus meeting, 27–28 Sept 2010, Geneva
61. Snow RW, Guerra CA, Mutheu JJ, Hay SI (2008) International funding for malaria control in relation to populations at risk of stable *Plasmodium falciparum* transmission. *PLoS Med* 5: e142
62. Price RN, Tjitra E, Guerra CA et al (2007) Vivax malaria: neglected and not benign. *Am J Trop Med Hyg* 77(suppl 6):79–87
63. Baird JK (2007) Neglect of vivax malaria. *Parasitol Today* 23:533–539
64. Harrison G (1978) Mosquitoes, malaria and man: a history of the hostilities since 1880. John Murray, London, p 314
65. McCullough D (1978) The path between the seas: the creation of the Panama Canal 1870–1914. Simon & Schuster, New York, p 698
66. Takken W, Snellen WB, Verhave JP, Knols BGJ, Atmoswedjono S (1990) Environmental measures for malaria control in Indonesia – an historical review on species sanitation, Wageningen Agricultural University Papers 90-7, p 167
67. Killeen GF (2003) Following Soper's footsteps: northeast Brazil 63 years after eradication of *Anopheles gambiae*. *Lancet Infect Dis* 3:663–666
68. Coatney GR (1963) Pitfalls in a discovery: the chronicle of chloroquine. *Am J Trop Med Hyg* 12:121–128
69. Sharma VP (1996) Re-emergence of malaria in India. *Indian J Med Res* 103:26–45
70. Baird JK (2000) Resurgent malaria at the millennium: control strategies in crisis. *Drugs* 59:719–743
71. Gianessi LP, Puffer CA (1992) Reregistration of minor pesticides: some observations and implications. In: USDA (ed) Inputs situation and outlook report. USDA, Washington, DC, pp 52–60
72. Packard RM (1998) 'No other logical choice': global malaria eradication and the politics of international health in the post-war era. *Parassitologica* 40:217–229
73. WHO (2006) Pesticides and their application for the control of vectors and pests of public health importance, 6th edn. Department of Control of Neglected Tropical Diseases, WHO Pesticide evaluation scheme (WHOPES), Geneva, p 125
74. Roberts DR, Manguin S, Mouchet J (2000) DDT house spraying and re-emerging malaria. *Lancet* 356:330–332
75. Roberts DR, Laughlin LL, Hsheit P, Legters LJ (1997) DDT, global strategies, and a malaria control crisis in South America. *Emerg Inf Dis* 3:295–302
76. Attaran A, Maharaj R (2000) Ethical debate: doctoring malaria badly; the global campaign to ban DDT. *BMJ* 321:1403–1405
77. WHO (2007) Position statement: the use of DDT in malaria vector control. Global Malaria Program, Geneva
78. N'Guessan R, Boko P, Odjo A, Chabi J, Akogbeto M, Rowland M (2010) Control of pyrethroid and DDT-resistant *Anopheles gambiae* by application of indoor residual spraying or mosquito nets treated with a long-lasting organophosphate insecticide, chlorpyrifos-methyl. *Malar J* 9:44
79. Protopopoff N, Verhaeghen K, Van Borel W et al (2008) A significant increase in *kdr* in *Anopheles gambiae* is associated with an intensive vector control intervention in Burundi highlands. *Trop Med Int Health* 13:1479–1487
80. Czeher C, Labbo R, Arzika I, Duchemin JB (2008) Evidence of increasing Leu-Phe knockdown resistance mutation in *Anopheles gambiae* from Niger following a nationwide long-lasting insecticide-treated nets implementation. *Malar J* 7:189
81. Grieco JP, Achee NL, Chareonviriyaphap T et al (2007) A new classification system for the actions of IRS chemicals traditionally used for malaria control. *PLoS One* 2(8):156–167
82. The Global Fund to Fight AIDS, Tuberculosis and Malaria. Global Fund Policy on Quality Assurance for Pharmaceutical Products: procurement of single and limited source pharmaceuticals. Tenth Board Meeting, Geneva, 21–22 Apr 2005, GF-B10-9, Annex 4, p 47
83. Carrara VI, Zwang J, Ashley EA et al (2009) Changes in the treatment responses to artesunate-mefloquine on the northwestern border of Thailand during 13 years of continuous deployment. *PLoS One* 4:e4551
84. Dundorp AM, Newton PN, Mayxay M et al (2004) Fake antimalarials in Southeast Asia are a major impediment to malaria control: multinational cross-sectional survey on the prevalence of fake antimalarials. *Trop Med Int Health* 9:1241–1246
85. White NJ (2010) Artemisinin resistance – the clock is ticking. *Lancet* 376:2051–2052
86. Field JW (1938) Notes on the chemotherapy of malaria. Bulletin from the Institute for Medical Research, Federated Malay States 2:63–81
87. Fiammetta R (2003) The miraculous fever tree: malaria and the quest for a cure that changed the world. Harper Collins, New York
88. Joy RJT (1999) Malaria in American troops in the South and Southwest Pacific in World War II. *Med Hist* 43:192–207
89. Downs WG, Harper PA, Lisansky ET (1947) Malaria and other insect-borne diseases in the South Pacific Campaign, 1942–1945. II. Epidemiology of insect-borne diseases in Army troops. *Am J Trop Med* 27:69–89
90. Shannon JA (1946) Chemotherapy of malaria. *Bull NY Acad Med* 22:345–357

91. Slater LB (2009) War and disease: biomedical research on malaria in the 20th century (Critical issues in health and medicine). Rutgers University Press, New Brunswick, p 272
92. Brosius OT (1927) Section II. Plasmochin in malaria. In: Sixteenth Annual Report. Medical Department, United Fruit Company, Boston, pp 26–81
93. Board for the Coordination of Malaria Studies (1943) Results of study of plasmochin toxicity: cases occurring in large scale use of the drug for suppression and treatment of malaria. Malaria Report No.290. Archive of the National Academy of Sciences, Washington, DC
94. Board for the Coordination of Malaria Studies (1943) Study of the curative action of plasmochin in vivax malaria: studies in Army and Navy installations. Malaria Report No. 358. Archive of the National Academy of Sciences, Washington, DC
95. Office of the Surgeon General (1943) The drug treatment of malaria, suppressive and clinical. Circular letter No. 153. JAMA 123:205–208
96. Carson PE, Flanagan CL, Ickes CE, Alving AS (1956) Enzymatic deficiency in primaquine-sensitive erythrocytes. Science 14:103–139
97. Most H et al (1946) Chloroquine for treatment of acute attacks of vivax malaria. JAMA 131:963–967
98. Schmidt LH et al (1977) Radical cure of infections with *Plasmodium cynomolgi*: a function of total 8-aminoquinoline dose. Am J Trop Med Hyg 26:1116–1128
99. Burgess RW, Bray RS (1961) The effect of a single dose of primaquine on gametocytes, gametogony and sporogony of *Laverania falciparum*. Bull World Health Organ 24:451–456
100. Rieckmann KH, McNamara JV, Frischer H, Stockert TA, Carson PE, Powell RD (1968) Gametocytocidal and sporontocidal effects of primaquine and of sulfadiazine with pyrimethamine in a chloroquine-resistant strain of *Plasmodium falciparum*. Bull World Health Organ 38:625–632
101. Peters W (1987) Chemotherapy and drug resistance in malaria, 2nd edn. Academic, London, p 1100
102. Ockenhouse CF, Magill A, Smith D, Milhous W (2005) History of US military contributions to the study of malaria. Mil Med 170(Suppl 1):12–16
103. Oyediran AB, Heisler M (1999) Malarone donation program. J Travel Med 6(Suppl 1):28–30
104. Korsinczyk M, Chen N, Kotecka B, Saul A, Rieckmann K, Chen Q (2000) Mutations in *Plasmodium falciparum* cytochrome b that are associated with atovaquone resistance are located at a putative drug-binding site. Antimicrob Agents Chemother 44:2100–2108
105. Ringwald P, Basco LK (1998) Malarone-donation programme in Africa. Lancet 351:673–674
106. Oyediran AB, Ddumba EM, Ochola SA, Lucas AO, Koporc K, Dowdle WR (2002) A private-public partnership for malaria control: lessons learned from the Malarone Donation Programme. Bull World Health Organ 80:817–821
107. Shretta R, Walt G, Brugh R, Snow RW (2001) A political analysis of corporate drug donation: the example of Malarone in Kenya. Health Policy Plan 16:161–170
108. Attaran A, Barnes KI, Curtis C et al (2004) WHO, the Global Fund, and medical malpractice in malaria treatment. Lancet 363:237–240
109. Cappelini MD, Fiorelli G (2008) Glucose-6-phosphate dehydrogenase deficiency. Lancet 371:64–74
110. Alving AS, Johnson CF, Tarlov AR, Brewer GJ, Kellermeier RW, Carso PE (1960) Mitigation of the hemolytic effect of primaquine and enhancement of its action against exoerythrocytic forms of the Chesson strain of *Plasmodium vivax* by intermittent regimens of drug administration. Bull World Health Organ 22:621–631
111. Crockett M, Kain KC (2007) Tafenoquine: a promising new antimalarial agent. Expert Opin Investig Drugs 16:705–715
112. Tinley KE, Loughlin AM, Jepson A, Barnett ED (2010) Evaluation of a qualitative enzyme chromatographic test for glucose-6-phosphate dehydrogenase deficiency. Am J Trop Med Hyg 82:210–214
113. Khim S, Nguon C, Guillard B, Socheat D, Chy S, Sum S, Nhem S, Bouchier C, Tichit M, Christophel E, Taylor R, Baird JK, Menard D (2011) Performance of the CareStart G6PD deficiency screening test. PLoS One (in press)
114. Pannaciuoli I, Tizianello A, Ajmar F, Salvadio E (1965) The course of experimentally induced hemolytic anemia in a primaquine-sensitive Caucasian: a case study. Blood 25:92–95
115. George GN, Sears DA, McCurdy PR, Conrad ME (1967) Primaquine sensitivity in Caucasians: hemolytic reactions induced by primaquine in G6PD deficient subjects. J Lab Clin Med 70:80–93
116. Baird JK, Surjadja S (2011) Consideration of ethics in primaquine therapy against malaria transmission. Trends Parasitol 27:11–16
117. Shekalaghe S, Drakely C, Gosling R et al (2007) Primaquine clears submicroscopic *Plasmodium falciparum* gametocytes that persist after treatment with sulphadoxine-pyrimethamine and artesunate. PLoS One 2:e1023
118. Shekalaghe SA, ter Braak R, Daou M et al (2010) In Tanzania, hemolysis after a single dose of primaquine coadministered with an artemisinin is not restricted to glucose-6-phosphate dehydrogenase deficient (G6PD A-) individuals. Antimicrob Agents Chemother 54:1762–1768
119. Butcher GA (1997) Antimalarial drugs and the mosquito transmission of *Plasmodium*. Int J Parasitol 27:975–987
120. Imwong M, Snounou G, Purkittayakamee S et al (2007) Relapses of *Plasmodium vivax* infection usually result from activation of heterologous hypnozoites. J Infect Dis 195:927–933
121. Baird JK, Hoffman SL (2004) Primaquine therapy for malaria. Clin Infect Dis 39:1659–1667
122. Coggeshall LT (1952) The treatment of malaria. Am J Trop Med Hyg 1:124–131
123. Trager W, Jensen JB (1976) Human parasites in continuous culture. Science 193:673–675
124. Malaria Vaccine Initiative (2011) Fact sheet: the RTS, S malaria vaccine candidate, Mar 2011, Washington, DC

125. Sherman IW (2009) The elusive malaria vaccine: miracle or mirage? American Society for Microbiology, Washington, DC, p 391
126. Committee on US Military Malaria Vaccine Research (2006) Battling malaria: strengthening the U.S. military vaccine program. National Academies Press, Washington, DC, p 144
127. Desowitz RS (1993) The malaria capers: tales of parasites and people. W.W. Norton, New York, p 288
128. Doolan DL, Dobano C, Baird JK (2009) Acquired immunity to malaria. Clin Microbiol Rev 22:13–36
129. Hviid L (2005) Naturally acquired immunity to *Plasmodium falciparum* in Africa. Acta Trop 95:270–275
130. Marsh K, Forster D, Waruiru C et al (1995) Indicators of life-threatening malaria in African children. N Engl J Med 332:334–336
131. Crompton PD, Pierce SK, Miller LH (2010) Advances and challenges in malaria vaccine development. J Clin Invest 120:4168–4178
132. Graves PM, Gelband H (2009) Vaccines for preventing malaria (SPf66). Cochrane Libr 2:1–30
133. Carrington WJ, Deragiache E (1999) How extensive is the brain drain? Finance Dev 36:46–49
134. Carpenter KJ (2000) Beriberi, white rice and vitamin B: a disease, a cause and a cure. University of California Press, Berkeley, p 296
135. Post P, Frederick WH, Heidebrink I, Sato S (eds) (2009) The encyclopedia of Indonesia in World War II. Brill, Leiden, Netherlands, p 710
136. Stone R (2010) Righting a 65-year-old wrong. Science 329:30–31

Books and Reviews

- Feacham GA, Phillips AA, Targett GA (eds) (2009) Shrinking the malaria map: a prospectus on malaria elimination. University of California San Francisco, Global Health Group, San Francisco, p 187
- Lobo L (2010) Malaria in the social context: a study in western India. Routledge, New Delhi, p 213
- Packard RM (2007) The making of a tropical disease: a short history of malaria. The Johns Hopkins University Press, Baltimore, p 320
- Shah S (2011) The fever: how malaria has ruled humankind for 500,000 years. Picador, New York, p 320
- Sherman IW (2010) Magic bullets to conquer malaria: from quinine to qinghaosu. American Society for Microbiology Press, Washington, DC, p 298
- Snowden F (2006) The conquest of malaria: Italy, 1900–1962. Yale University Press, New Haven, p 304
- Winther PC (2005) Anglo-European science and the rhetoric of empire: malaria, opium and British rule in India 1756–1895. Lexington Books, New York, p 450
- Yip K-C (ed) (2009) Disease, colonialism and the state: malaria in modern East Asian history. Hong Kong University Press, Hong Kong, p 161

True Color Night Vision Video Systems in Intelligent Vehicles

DIRK HERTEL
Quincy, MA, USA

Article Outline

Glossary
Definition of the Subject and Its Importance
Introduction
The Automotive Imaging Application
Human Perception of the Automotive Imaging Environment
The Automotive Imaging Environment
Extending the Dynamic Range
Extending Spectral Sensitivity into NIR
Adding Color Differentiation to VIS + NIR-Sensitive Imager
Visual Representation of the VIS + NIR Color Information
Conclusion
Future Directions
Bibliography

Glossary

Automotive night vision A technical system to enhance human night vision in situations where visibility is limited or challenging to the human visual system, for example in situations where the range of distances illuminated by headlights is shorter than the safe braking distance or in situations where the human eye may be too slow to adapt to sudden changes in brightness. It requires an effective human–machine interface, for example, a visual display to make available to the driver real-time additional visual information.

CMOS image sensor An active-pixel image sensor (APS) consisting of an integrated circuit with an array of light-sensitive cells (pixels) and manufactured using complementary metal–oxide–semiconductor (CMOS) technology. Each pixel combines a photodetector with an active amplifier that may perform additional functions such as controlling the pixel response to light and extending the pixel's dynamic range.

Driver assistance system (DAS) A technical system to assist the driver in the task of driving, especially in situations where the human sensory system (visual, auditory) or human attention might fail to recognize a dangerous situation, or when the human response time in the perceptual-motor task of recognizing a dangerous situation then using vehicle controls to take corrective action may be too slow to avoid an accident.

Dynamic range Dynamic range is the ratio between the largest and smallest possible signal values of a changeable quantity such as light. For image sensors and cameras, dynamic range refers to input radiance or luminance signals and is defined as the ratio between the maximum and minimum signals where a camera can simultaneously capture detail carried by small signal variations.

Electromagnetic spectrum, visible (VIS) and near-infrared (NIR) The human eye can detect the visible (VIS) portion of the electromagnetic spectrum ranging from wavelengths of about 380 to 730 nm. Infrared is electromagnetic radiation with longer wavelengths than visual, with NIR denoting the shortest wavelengths of the IR-spectrum between 700 and 1,400 nm.

Wide dynamic range (WDR) Also referred to as high dynamic range (HDR), it covers a family of techniques to extend the dynamic range of image sensors beyond that of an image sensor with a linear relationship between input radiance and digital output signal.

Definition of the Subject and Its Importance

Vision-based driver assistance aims to increase traffic safety at night by enhancing human night vision in situations where visibility is limited or challenging to the human visual system. The night vision cameras described sense the near-infrared radiation of the car's own headlights to extend the range of forward vision, while, at the same time, discriminating traffic-relevant colors of signal lights and signage. Adaptive wide-dynamic-range image sensor technology ensures that scene detail in both deep shadow and bright highlight is captured even in situations where the human eye may be blinded or too slow to adapt to sudden changes in brightness. The information from the near-infrared

and visible spectrum is combined into natural-looking color images and communicated to the driver via on-board visual displays to enable recognition of the traffic situation.

Introduction

Since the late 1990s, there has been research into driver assistance systems employing sensors to perceive the automotive environment and utilize their outputs to create warning and vehicle control signals. Vision-based systems are of particular importance because more than 90% of driving decisions and actions are based on vision. Although the human visual system (HVS) is able to adapt to a vast range of lighting conditions, the rate of visual adaptation is often outpaced by the rate of change in lighting conditions while driving, for example when passing blinding headlights at night. Camera-based vision systems have the capability to aid the driver in critical situations by more rapidly adapting to changes in brightness and intra-scene dynamic range. They may also extend spectral sensitivity into the NIR to utilize the invisible portion of the headlight spectrum or can additionally employ NIR-only headlights that can be used in high-beam mode without blinding other drivers. Since the visual channel has the highest capacity of information transfer compared to other senses, image displays are used as human-machine interface to communicate information to the driver. However, the displayed image must be of sufficient quality to meet minimum requirements of usefulness so that scene objects can be seen, and of naturalness so that scene objects can be instantly recognized. For example, displayed objects with too little contrast will be overlooked while objects with unnatural grayscale and color appearance or distorted perspective may not be recognized in a timely fashion.

Although the majority of research on driver assistance systems focuses on image processing to interpret the information delivered by image sensors, comparatively little attention is given to the task of capturing scene detail. Since the camera forms the first link in the image information processing chain, its performance is critical; if the intra-scene dynamic range, the ratio of the brightest to darkest information-carrying intensity, exceeds the dynamic range of the

image sensor, then scene detail will be lost and cannot be recovered by image processing.

A model-based methodology is one that:

- Characterizes the automotive imaging environment (photospace) as wide dynamic range (WDR)
- Defines incremental signal-to-noise ratio (iSNR) as the criterion for reliably capturing WDR scene detail
- Uses the inherent sensitivity of silicon-based image sensors to utilize scene information carried by NIR radiation
- Separates color and NIR information in the image sensor and WDR image signal processing
- Defines the principles for combining luminance information from the visual and NIR spectral ranges with traffic-relevant color information into a natural and useful image for visual display

The Automotive Imaging Application

Reduced visibility at night substantially increases the risk of traffic accidents caused by misjudgment of dark roadways or failure to recognize the presence of pedestrians and obstacles. If an accident does occur, passive safety systems such as seat belts and airbags can reduce the percentage of fatalities to the vehicle's occupant, but only to a limited degree [1]. Any further reduction of traffic fatalities requires the use of active driver assistance systems.

Driver Assistance

Driver assistance can be classified into comfort systems and safety systems. These systems employ sensors that take measurements of the automotive environment, interpret the signals delivered by the sensors, generate warning signals, and, if the driver takes no action, even take over vehicle controls. The distinction between comfort and safety is fluid; parking guidance makes parallel parking more comfortable but also enhances safety by detecting persons and objects not visible to the driver. As an example of active safety, lane-keeping systems analyze sensor signals to determine if the vehicle is leaving the lane or roadway, then generate warning signals and in case of driver inaction steering signals. Other examples are lane-change assistance with blind-spot monitoring, collision avoidance, and driver-drowsiness monitoring.

Vision-Based Driver Assistance Vision-based driver assistance has a high potential of increasing traffic safety. First, more than 90% of the information that is critical for making driving decisions comes through the visual channel. Second, conditions that impair the flow of visual information have a significant impact on traffic safety. Nighttime driving is twice as risky as daytime driving, even though the volume of traffic is often less at night. Nighttime accidents tend to be more severe, with a disproportionate increase of fatalities compared to the volume of traffic. Nighttime accidents more often involve pedestrians or cyclists. The illumination of the roadway may be inadequate. For example, dipped headlights stretch over a fixed range of about 180 ft, but the safe braking distance increases with speed exceeding the visibility range at speeds greater than 30 mph [2].

Video-based driver assistance systems provide additional image information in situations where visibility is poor (night vision) or obstructed (side or rear vision). Applications range from driver assistance with forward and rear view images displayed on a dashboard screen to active safety systems where the video footage is analyzed by software that can issue warnings or trigger vehicle controls. On the front-end of these systems are video cameras that must capture scene details reliably under all circumstances.

Requirements of Vision Systems for Driver Assistance The basic requirements for a vision-based driver assistance system are image quality and reliability.

Image Quality Image quality can be defined as “the degree to which the image can be successfully exploited by the observer” [3]. An image of “high” quality should satisfy three criteria: genuineness, usefulness, and naturalness (GUN-model).

- Genuineness is described as the degree of apparent similarity of the reproduced image and its viewing environment with the external reference, that is, the original image and environment. Genuineness is related to fidelity, for example between a painting and its photographic reproduction.
- Usefulness is how relevant the reproduced image and environment are to the observer and the task activity. Usefulness is related to the discriminability of objects in the image, for example obstacles on a dark roadway.

- Naturalness is described as the degree of apparent similarity between the reproduced image and environment and the viewer's internal references, that is, memory prototypes. It is related to recognizability of objects in the image, for example whether the obstacles on a dark roadway are pedestrians or deer.

Night vision systems for driver assistance do not require genuineness, but they do require both usefulness and naturalness. Color and near-infrared (NIR) can supply important additional information that increases the usefulness of a night vision image, though NIR sensitivity may affect color saturation and thus reduce the naturalness of displayed images (section “[Adding Color Differentiation to VIS + NIR-Sensitive Imager](#)”).

Image information that is interpreted by software has to fulfill the criterion of usefulness. Image information displayed for viewing has to be natural so that it can be interpreted instantly. An unnatural image requires effort by the interpreter and thus can distract, for example, false colors, inverted grayscale (thermal images), or distorted perspective (extreme fisheye).

Reliability The main difference between photography and automotive imaging is that the former is in most cases controlled by a conscious subject – the photographer – who selects the frame, knows what is in the scene, and can avoid extremes (strong light sources). However, when a camera is fixed in a vehicle, frame selection and content are arbitrarily determined by vehicle position, so intra-scene dynamic range may include, for example, approaching headlights, glare from other vehicles, tunnel entrances or exits, or solar glare. Image detail provides information about the objects in the real world, such as their location, shape, size, brightness, and color. If the dynamic range of a camera is too narrow to accommodate the intra-scene dynamic range, important object details will be missing from the corresponding image. Once lost by capture failure, these details can never be recreated with image post-processing [4, 5].

Conclusion: Model-Based Approach The design of automotive vision systems is governed by performance versus cost trade-offs. Therefore, it is important that the design is based on a clear understanding of

the nonnegotiable performance thresholds such as sensitivity, signal-to-noise ratio, and dynamic range. A model-based design can quantitatively define and meet the performance requirements of automotive night vision. This includes models of human color night vision, the automotive night vision environment (photospace), and detection of scene detail (signal-to-noise ratio).

Human Perception of the Automotive Imaging Environment

The human visual system (HVS) is able to adapt to a vast range of natural and artificial lighting conditions. The mechanisms of adaptation are complex and include the combined responses of iris, lens, retinal photoreceptors, plus processing in the retina and visual cortex.

Day and Night Vision

Viewing the environment from a fast moving vehicle means that the driver's visual system must adapt to a visual environment that changes more rapidly than for a pedestrian or stationary observer. Under daylight conditions, such changes requiring instant adaptation include driving into the deep shadow of a bridge or tunnel, emerging from deep shadow into bright daylight, and solar glare. At night, the most challenging situation occurs immediately after passing the blindingly bright headlights of oncoming vehicles and having to adapt to the comparatively dark roadway ahead. These examples illustrate that adaptation of the HVS to instantly changing stimuli is the most critical issue in the automotive environment.

However, the models describing human day and night vision refer to steady states of global adaptation to defined ranges of light levels and characterize the responses of the two types of retinal sensors, rods and cones, to these light levels. Three distinct adaptation states of the human eye have been identified: photopic, mesopic, and scotopic.

- Photopic vision in bright light is dominated by cone receptors of relatively low sensitivity that provide full color at high acuity, which is highest at a central foveal field of about 2°. The spectral sensitivity function for photopic vision is measured as the

combined response of the long-, medium-, and short-wavelength-sensitive cones and standardized as the photopic luminous efficiency function $V(\lambda)$ with peak sensitivity at 555 nm (yellowish green).

- Scotopic vision, at very low light levels, relies on highly sensitive rods that provide only a monochrome image with poor acuity at the center of the field of vision but increased acuity toward the periphery. The peak of the scotopic luminous efficiency function $V'(\lambda)$ is at 510 nm, shifted toward bluish green, the so-called Purkinje-shift.
- Mesopic vision is a transition phase between photopic and scotopic vision that is invoked at luminance levels below 10 cd/m^2 and above 0.01 cd/m^2 . The mesopic luminous efficiency function has not yet been standardized. Proposals range from a simple weighted average of the photopic and scotopic sensitivity functions to more complex chromatic models that weigh the contributions from the individual cones and yield sensitivity functions with two or three peaks [6]. Mesopic vision thus represents a balance between photopic and scotopic vision where the response of monochrome and color receptors will depend not only on overall light level but also on the scene-dependent local distribution of brightness on the retina. For example, the response within those areas of the retina that are stimulated by bright colored traffic lights is still cone-dominated while the response of retinal areas that are stimulated by dark background objects is rod-dominated.

Due to the presence of artificial lights in the automotive environment at night, the eye's adaptation never quite reaches the scotopic state but remains between or in the photopic and mesopic states. Color vision in very dark scenes is restricted to bright colored lights (traffic, tail, and brake lights), retroreflective signage, and surfaces brightly illuminated by headlights and streetlights. As these are relatively small and isolated by a dark background, they are referred to as "unrelated colors" [7]. Thus, relevant traffic lights are perceived as colored, but most dark objects (foliage, pedestrian's clothes) are perceived as monochrome.

Although the simultaneous dynamic range of the human eye, also known as the instantaneous sensitivity window, is limited to about 10,000:1, adaptation

enables the HVS to adapt over time to a much wider range of luminance levels (Fig. 5). The rate of adaptation of the HVS to changing light levels is critical to traffic safety. Adaptation works with two time constants: fast (transient) adaptation occurs within seconds whereas slow adaptation happens on a much longer timescale, ranging from several minutes up to three quarters of an hour. Fast adaptation is the instant reaction to a sudden change in brightness, for example tunnel entry or exit, and vision is impaired until adaptation is achieved, usually after 1–2 s. Therefore, fast adaptation is of utmost importance in the performance of the HVS during driving. The rate of slow adaptation depends on the direction (dark takes longer than bright adaptation) and the light levels before and after adaptation. For example, the brighter the initial light level and the darker the final light level, the more time is required for slow dark adaptation. Headlight glare from oncoming vehicles constantly interrupts the process of dark adaptation.

Challenges to the Human Visual System

Driving under night conditions is very stressful to the HVS. First, the range of luminance levels is extremely large, ranging from blinding oncoming headlights to dark objects beyond the cone of one's own headlights. The constant change in peak brightness and position of the highlights puts the eye under the constant strain of global and local adaptation so that the retina is constantly registering either photopic or mesopic brightness levels. Color vision is restricted to bright lights and brightly illuminated objects whose position in the field of vision changes rapidly and constantly. The rate of change in stimuli will almost always exceed the rate of transient (fast) adaptation. Visual night driving performance can be evaluated by the responses to three visual subtasks that are defined by questions and evaluated by corresponding parameters: "Can it be seen?" (contrast threshold), "How quickly?" (reaction time), and "What is it?" (recognition threshold) [6]. The ability of the HVS to fulfill these tasks decreases with age (Table 1). The most serious aspect of age-related visual deterioration is the slowing of transient adaptation, making it more likely that the driver is "blind" immediately after sudden changes of brightness. At night, headlight glare can lead to the

phenomenon of “black holes,” a temporary inability to perceive dark background detail after blinding. Although all drivers experience black holes, older

drivers take longer to recover vision after the experience. Many societies have an increasing proportion of older drivers, so the implications of this issue are ever greater.

The simulation in Fig. 1 illustrates the loss of sensitivity and shift in color vision due to yellowing with age.

True Color Night Vision Video Systems in Intelligent Vehicles. Table 1 Visual impact of aging on human vision performance

Parameter	Effect of aging	Visual impact
Sensitivity	Decreases	Less light reaches retina
Color perception	Deteriorates	Loss of blue sensitivity
Rate of dark adaptation	Slows	“Black holes”
Glare	Increases	Loss of dynamic range
Contrast sensitivity	Decreases	Loss of detail
Peripheral vision	Decreases	Restricted

Conclusion: Producing Natural Images Under Mesopic Conditions

Night vision images displayed for the driver have to be rendered “naturally” so that traffic-relevant objects can be quickly seen and recognized. This requires that illuminated dark details such as pedestrians are rendered as a sufficiently bright monochrome image while “unrelated color” objects such as traffic lights are rendered with their correct hue and sufficient saturation (Fig. 2). In other words, the best compromise is to reconstruct a good luminance image then add those



True Color Night Vision Video Systems in Intelligent Vehicles. Figure 1 Simulated change in night color vision with age (a) 20, (b) 60, and (c) 70 [42]



True Color Night Vision Video Systems in Intelligent Vehicles. Figure 2
Same scene showing (a) related color at daytime and (b) unrelated color at nighttime [8]

unrelated colors that are relevant to the traffic scene. This way, the automotive vision system will mimic the mesopic state of the HVS, but with two key improvements. First, as adaptation to sudden changes in light levels is almost instantaneous, the driver is provided with visual information by the safety system in situations where the HVS adapts slowly. Second, the displayed luminance image is brighter and more detailed at a wider field of view in comparison to the corresponding rod-based HVS perception of the dark background. Although color cameras do not experience a shift in spectral sensitivity akin to the Purkinje shift, image processing algorithms (such as white balancing) still have to be applied to correct color shifts in illuminated objects caused by altered illuminant color temperature, for example moving between areas lit by mercury and sodium vapor lamps.

The Automotive Imaging Environment

A possible methodology for estimating the dynamic range of the automotive vision environment is based on photometric and spectral measurements together with the statistical photospace model. In addition, knowledge of the spectral characteristics of colored light sources is important when designing VIS–NIR color night vision systems.

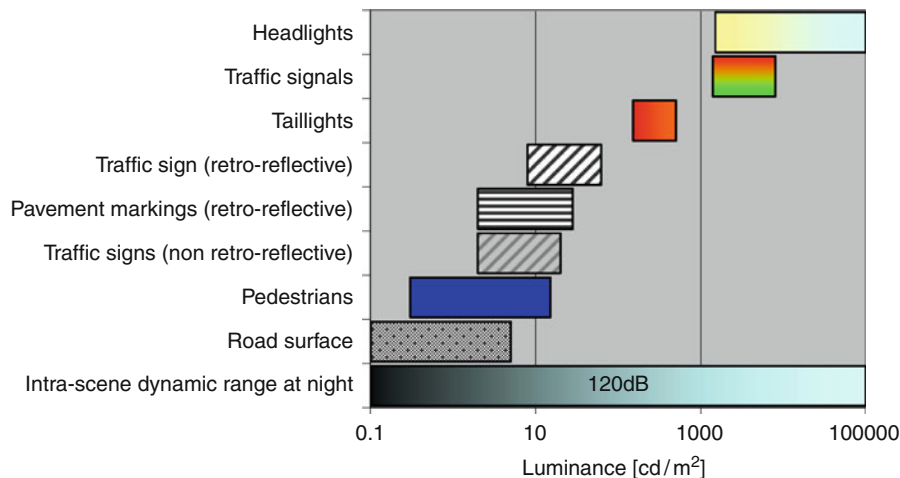
Scene Photometry

To determine dynamic range requirements, a quantitative model of the image capture environment is

needed. First, a measure of the brightness of light sources and illuminated objects in automotive scenes is required. The intensity of light is measured in radiometric power units, which are then converted to photometric units that represent the perceived brightness. This is done by weighting the radiant power at each wavelength by a spectral sensitivity function that models human brightness sensitivity, for example the photopic luminous efficiency function $V(\lambda)$ for the bright-adapted eye, or the scotopic luminosity function $V'(\lambda)$ for the dark-adapted eye. The amount of visible light reflected by or emitted from a surface area is measured as luminance with units of candela per square meter [cd/m^2]. Luminance is a good indicator of how bright a light source or illuminated object will appear. It measures the luminous power perceived at the surface of a light source or an illuminated object. It is independent of viewing distance and not changed by optical systems, e.g., lenses. Luminance measurements of objects in automotive scenes require spot luminance meters.

In the case of color and multispectral imaging, radiance must be measured [$\text{W}/\text{cm}^2/\text{sr}$] at specific wavelengths to characterize the night vision environment. This is done by a spot spectroradiometer that can measure not only the visible but also the NIR range.

Scene analysis relies on measuring the luminance of individual scene elements rather than the scene average [9]. Scene photometry in combination with scene analysis has been used to build statistical distributions



True Color Night Vision Video Systems in Intelligent Vehicles. Figure 3
Luminance ranges for light sources and reflective surfaces in typical night traffic scenes [8]

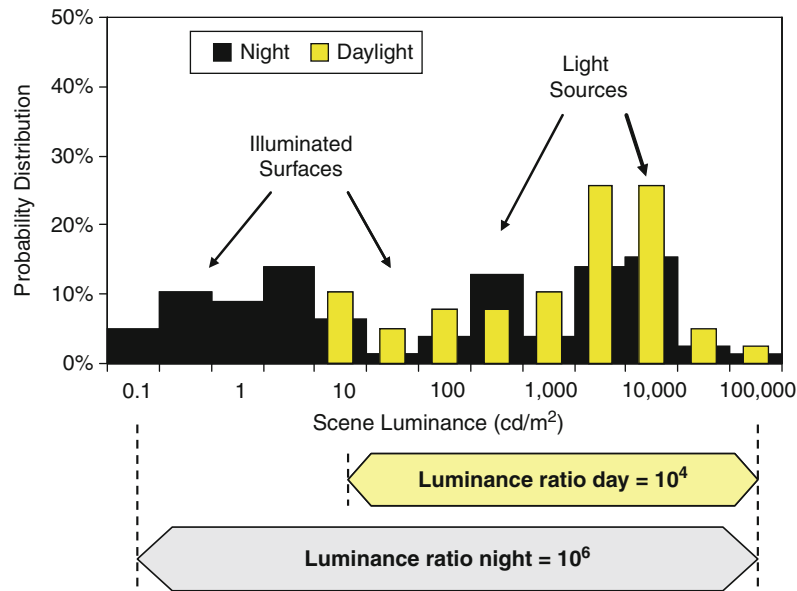
of light levels within a scene, ranging from the brightest lights to the darkest illuminated surfaces. Figure 3 shows an automotive vision example with luminance ranges for light sources and reflective surfaces typical of a night traffic scene.

Photospace in Automotive Vision

Photospace describes the environment as a statistical distribution that is populated by measuring average scene luminance levels for a great number of application-typical scenes [10]. By analyzing at least ten daytime and ten night traffic scenes, extended photospace distributions can be built based not on scene averages but rather on the probability that a scene element of a given luminance will occur (Fig. 4) [4, 11]. As more image data are added, such distributions will ever more closely represent application conditions. These extended photospace distributions for automotive vision by day and night show clusters representing light sources at high luminance levels (reflections of sun, vehicle lights, street lights, and traffic lights), and illuminated surfaces at lower luminance levels (road surface, traffic signs, and traffic participants). The luminance ratio within night scenes (10^6) is anticipated to be higher than that of daylight scenes (10^4 – 10^5) since poorly illuminated objects at night have much lower luminance. The highest luminance ratios in daylight

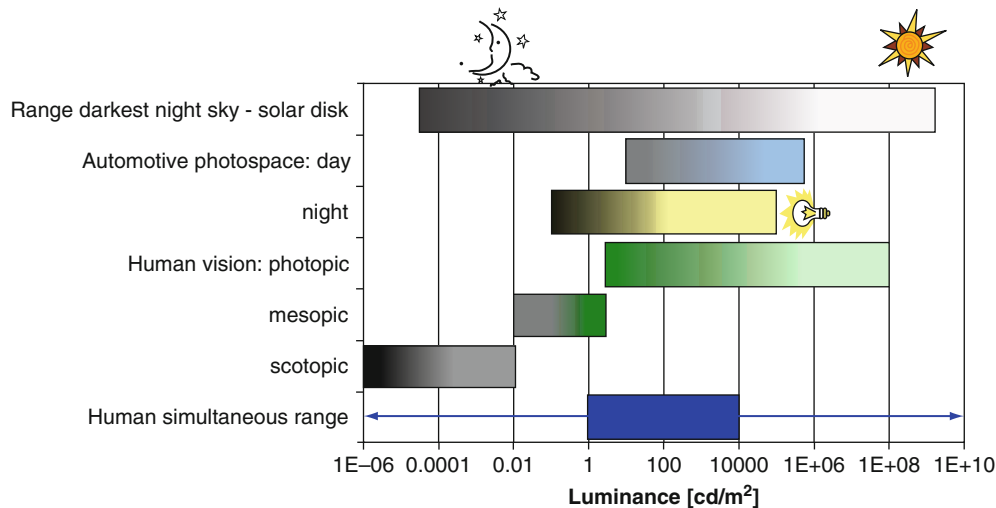
scenes occur when the roadway enters or exits deep shade (bridges, tunnels). For the camera to detect image detail at night simultaneously in both the brightest (differentiation of headlights) and darkest scene areas (pedestrian detection), a dynamic range of $10^6:1$ is required. However, a wide dynamic range is not the only requirement for an automotive camera to work in a night vision environment. The other necessity is that the camera is sensitive to the lowest luminance levels of the automotive photospace. The noise-limited sensitivity of the camera at the video frame rate should be such that its dynamic range covers the entire automotive photospace, including most of the mesopic range plus as much of the photopic range as possible (Fig. 5).

For a monochrome system, it is sufficient to render the brightest lights near the maximum signal level of the pixel in order to minimize halos and flare, thus maximizing differentiation between various lights, for example, headlights. However, in the case of color capture, this may be insufficient; a colored light captured near the maximum signal level will be too desaturated to be recognizable. As a rule of thumb, if a pixel's set of RGB signals is to be perceived as "green" (10% color saturation between the hue angles of 85° and 155°), each of the R and B signals has to be at least 10% lower on average than the G signal. However, if the green light increases in brightness, the G signal will reach then exceed



True Color Night Vision Video Systems in Intelligent Vehicles. Figure 4

Extended photospace distribution showing clusters for light sources and illuminated surfaces [5]



True Color Night Vision Video Systems in Intelligent Vehicles. Figure 5

Luminance levels and adaptation of the HVS [42]

saturation level before R and B, thus decreasing the G/R and G/B ratios so that the green color is barely distinguishable from “white” (see Fig. 6). Therefore, the dynamic range of a night vision color system must be wide enough to prevent any of the RGB color signals from being clipped.

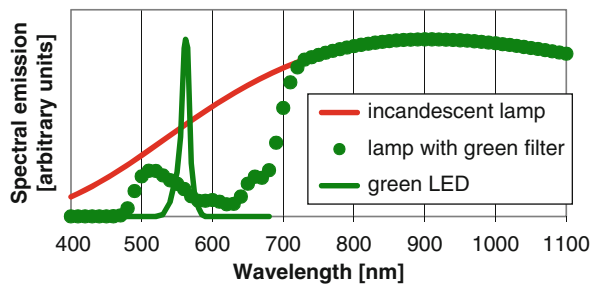
Spectral Distribution of Light Sources

The diversity of spectral characteristics of colored light sources poses a challenge to automotive vision systems that combine NIR sensitivity with color differentiation. For example, a green traffic light might be an LED array



True Color Night Vision Video Systems in Intelligent Vehicles. Figure 6

Typical urban night traffic color scene where limited dynamic range results in clipped highlight colors appearing *white* (centers of *green* traffic lights, *red* tail lights) [35]



True Color Night Vision Video Systems in Intelligent Vehicles. Figure 7

Spectral VIS–NIR emission curves of an incandescent lamp (3,200 K), an incandescent lamp with a green glass filter, and a green LED [35]

or an incandescent lamp behind a colored filter. The former has a narrow spectral distribution peaking in the 500–600 nm range. The latter has a more complicated spectral distribution. Although in the VIS range, it has a peak in the green region of the visible spectrum, since most commonly used color filter materials are almost transparent in the NIR range and the spectral emission of incandescent lamps (color temperature 3,200 K) peaks in the NIR at about 800 nm, the strong NIR contribution might exceed that from the VIS range (400–700 nm), see Fig. 7.

Spectral measurements of objects in automotive scenes require spot spectroradiometers.

This high NIR content of colored lights does not pose a problem for human vision and common RGB color cameras because the eye is insensitive to IR, and the latter has an IR-cut filter.

It follows that a VIS–NIR color night vision system has to combine high sensitivity in both spectral areas with good reproduction of those colors relevant to night traffic scenes. This requires the imager to have WDR to avoid signal saturation in colored lights, plus sufficient separation of NIR and color signals so that colors can be reproduced independently of the light source's IR content.

Conclusion: Automotive WDR Environment

Intra-scene dynamic range is determined by the range of intensities between bright lights (some of them carrying color information) and dark, dimly illuminated objects. The automotive vision environment can be measured by scene photometry to quantify the luminance ranges of traffic-relevant light sources and illuminated objects, and to populate statistical photospace distributions that show the probability that scene elements of a given brightness (luminance) will occur at night or day. Automotive scenes typically cover a WDR, highest at night due to lower background luminance levels. Challenges in daylight conditions are deep shadows (parked vehicle under bridge, tunnel entry) and blinding (solar glare, exit from tunnels or bridges). Night vision requires both high sensitivity (to distinguish shadow detail) and WDR (to distinguish highlight color detail). Preserving the color information of traffic-relevant lights requires an even wider dynamic range than for monochrome vision.

Extending the Dynamic Range

A possible methodology to extend the dynamic range of image sensors without compromising the transfer of scene information to image information can be developed from an information-theoretical model.

Definition of Dynamic Range

Dynamic range generally relates to an imager's ability to capture information about both bright and dark

objects in the same scene at the same exposure settings. There are several definitions of dynamic range. In its most general sense, dynamic range describes the ratio between the largest and smallest possible signal values of a changeable quantity such as light [4]. Dynamic range can be either input-referred (ratio of real-world light levels) or output-referred (ratio of the response signals delivered by a light detector). Since the output-referred definition can lead to dynamic range numbers built solely on technical specifications (such as the bit depth of an A/D converter) rather than on actual measurements, it has limited value in predicting dynamic range [13]. Therefore, dynamic range should refer strictly to input signals, for example luminance L in photopic units of cd/m^2 . The relationship between input and output signals is established from the imager response curve.

In silver halide photography, the luminance ratio of the highlights to the shadows is called “object ratio” and the logarithm of this ratio “object range” [4]. The equivalent of “object range” in digital imaging is the input-referred dynamic range, which can be expressed in units of decibels [dB] by calculating $10 \log$ of the luminance ratio. However, technical specifications of digital and video image sensors traditionally referred to ratios of output voltages or digitized light levels so that decibels represent $20 \log$ of the ratio (even when the represented input luminance is directly proportional to the output voltage or level) not to its square [14]. In order to avoid disadvantaging input-referred data, the $20 \log$ rule is also used to calculate the input-referred dynamic range from luminance ratios:

$$DR = 20 \cdot \log_{10} \left(\frac{L_{\max}}{L_{\min}} \right). \quad (1)$$

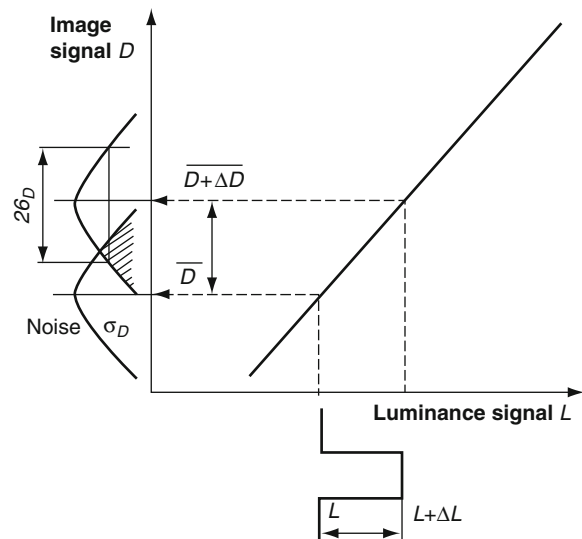
Reliability Criterion

Ascertaining dynamic range requires determining the maximum and minimum signal levels. Detection of the smallest possible signal is limited by the noise floor. To become detectable, a signal must exceed the noise floor; in other words, the signal-to-noise ratio (SNR) must be greater than 1. The largest possible detected signal is limited by saturation. Any input signal above the saturation level can only reproduce a saturation level output signal and thus remains indistinguishable from the saturation signal. This introduces the concept of signal

variation as the carrier of information: object detail is carried by small variations in light level. Scene information can only be gathered reliably if the image capture system can detect and reproduce the small signal variations carrying that information. Thus, dynamic range needs a more precise specification to ensure detection reliability: it is the ratio between the maximum and minimum luminance signals where a camera can simultaneously capture detail carried by small signal variations.

A distinction has to be made concerning the spatial dimension of object details. Below a certain size, the details in an image are no longer formed independently; the signal at a given image point depends not only on the luminance at the corresponding object point but also on that of its surroundings [4]. This effect can be accounted for by the modulation transfer function (MTF). However, this estimation of dynamic range refers only to the reproduction of details large enough that the loss of detail due to MTF can be ignored. It is desirable to deduce dynamic range from objective measurements. This methodology is laid down in a progression of ISO/TC42 imaging performance standards [15–17].

Figure 8 shows the transfer of large-area detail (small luminance variation) by the response curve in



True Color Night Vision Video Systems in Intelligent Vehicles. Figure 8

Transfer of large-area detail (small luminance variation ΔL) by the response curve $D(L)$ in the presence of noise σ_D [5]

the presence of noise [18]. The response curve is the relationship $D(L)$ between input luminance signal L and output digital image signal D . ISO/TC42 refers to the response curve as opto-electronic conversion function (OECF). The slope of the OECF, the change dD in image signal due to a change dL in the luminance signal, is referred to as incremental gain $g(L)$:

$$g(L) = \frac{dD}{dL}. \quad (2)$$

Detection error is given by noise. Since dynamic range refers to the input luminance signal L , the noise σ_L has to be defined as the luminance-equivalent of the noise σ_D that can be observed and measured in the digital image [19]. At each luminance level L , the equivalent luminance variation σ_L can be calculated from the image noise σ_D via the incremental gain:

$$\sigma_L(L) = \frac{\sigma_D(L)}{g(L)}. \quad (3)$$

The criterion for detection reliability is the SNR, the ratio of the input luminance signal L and the luminance-equivalent noise σ_L . Detection reliability requires the SNR to stay above a threshold of one to fulfill the following condition:

$$\frac{L}{\sigma_L(L)} = \frac{g(L) \cdot L}{\sigma_D(L)} > 1. \quad (4)$$

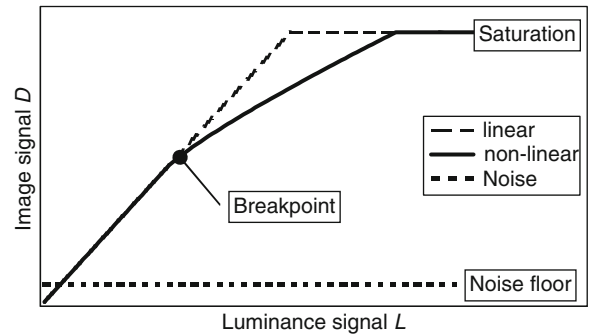
The expression in (4) is referred to as “incremental SNR” (iSNR, also S/N_x and SNR_{inc}); the incremental signal is defined as $iS(L) = g(L) \cdot L$. An equivalent concept for determining detection reliability in silver halide photography is known as noise-equivalent contrast (NEC) [20].

The threshold condition $iSNR > 1$ means that at any given luminance level L , detail carried by a small variation in luminance can only be reliably detected if the corresponding incremental image signal is high enough and the noise low enough. Fulfilling the iSNR criterion will be first explored for imagers with a linear response. At very low luminance levels, noise keeps iSNR below its threshold. The minimum reliable luminance level L_{min} is reached only when iSNR rises above the threshold. At very high luminance levels, the onset of saturation will clip further signal increases and thus cause the incremental gain to drop sharply. The maximum reliable luminance level L_{max} is reached

when iSNR once more drops below the threshold. It follows that dynamic range for linear imagers is bounded by the two extreme luminance levels L_{min} and L_{max} at which the iSNR condition of (4) still holds.

The iSNR condition is now applied to imagers with extended dynamic range. This requires shifting saturation to higher input signal levels by a nonlinear response curve: at a so-called breakpoint, the response curve’s slope (incremental gain) is suddenly lowered to delay saturation to higher luminance levels than for a linear imager (Fig. 9). However, lowering incremental gain will, in turn, decrease iSNR, pointing to a potential iSNR trade-off when increasing dynamic range. Rather than just using iSNR to determine the boundary luminance levels L_{min} and L_{max} , the minimum iSNR criterion has to be maintained at every luminance level in between. Verifying this requires iSNR be measured as a function of luminance L . An equivalent strategy was used for silver halide photography where NEC curves were measured as function of exposure [20].

Therefore, nonlinear imagers require a more general definition beyond the recommendations of ISO/TC42; dynamic range is the input luminance range within which the threshold criterion for $iSNR(L)$ is met at all levels $L_{min} \leq L \leq L_{max}$. This suggests a strategy for maximizing the dynamic range of imagers with a nonlinear response curve: maintain the necessary minimum of $iSNR(L)$ over the largest possible range of luminances L . Thus, iSNR becomes a metric describing the reliability of detecting scene detail. The theoretical detection limit is 1, but the reliability requirements of applications may require



True Color Night Vision Video Systems in Intelligent Vehicles. Figure 9

Extending dynamic range with a nonlinear response curve [5]

higher thresholds. For example, a minimum of 10 is suggested for “acceptable” photographic image quality, and between 1.5 and 4 are estimated for machine-vision applications [18].

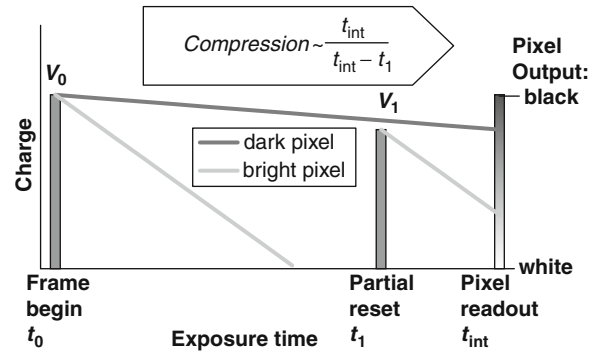
Technologies to Extend Dynamic Range

The following section reviews a selection of WDR technologies and explores the iSNR trade-offs of nonlinear dynamic range extension. A WDR video camera for automotive applications should combine good sensitivity with low noise at a given frame rate (10–60 fps) plus a WDR extension that can be adapted quickly to match typical automotive photospace conditions (80–120 dB). The first group of technologies for WDR extension are those that directly change the CMOS pixel response function (such as logarithmic response [21], linear-log combination [22], in-pixel light-to-frequency conversion [23], pixel-level A/D conversion [24], or pixels with overflow capacitors [25]), while a second group changes pixel response indirectly by altering the timing (multiple exposure, multiple partial reset [26]).

The first group, directly changing the response function, involves extensive pixel circuitry resulting in low fill-factor and thus poor low-light sensitivity. For example, in-pixel light-to-frequency conversion provides 115–130 dB WDR, but the required circuitry decreases the fill factor to 25%, and reduces low-light sensitivity. Similar sensitivity degradation currently limits the usefulness for automotive video of both pixel-level A/D conversion and pixels with overflow capacitors. Although pixels with direct logarithmic compression can provide WDR above 170 dB at low cost, the WDR extension is not adaptive, and at present, high fixed-pattern noise and dark current decrease low-light SNR. The combined linear-log CMOS pixel does avoid these drawbacks, but the required seven transistors per pixel limit the fill factor.

WDR can be covered with a linear image sensor by multiple exposure, but application to automotive video requires either fast readout (high frame rate with corresponding loss in low-light sensitivity), or local frame storage and processing.

Among WDR technologies that alter the timing, CMOS image sensors with multiple partial reset (multiple-slope) technology offer good imaging



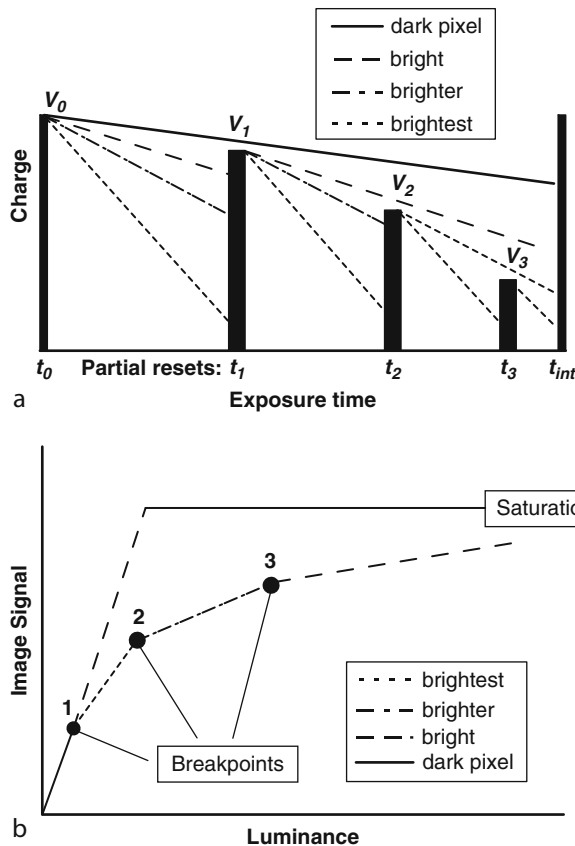
True Color Night Vision Video Systems in Intelligent Vehicles. Figure 10

Charge-time diagram for a CMOS WDR pixel with a single partial reset [5]

performance at low cost and are thus well suited to automotive video capture [4]. The example of a multiple-slope imager explains how to extend dynamic range while maintaining detection reliability.

At the heart of the CMOS pixel is a light-sensitive photodiode with a parasitic capacitance that is charged by applying a voltage pulse V_0 at frame begin t_0 , then discharged as exposure generates photoelectron-hole pairs. Figure 10 shows the charge-time diagram of a CMOS pixel. The rate of discharge depends on the input luminance; a darker pixel is discharged at a lower rate, a brighter at a higher one. After integration time t_{int} , the remaining charge is read out and converted into the image signal D . This is proportional to the degree of discharge between t_0 and t_{int} ; no discharge corresponds to $D = 0$, and full discharge to saturation level (Fig. 10). A darker pixel will only be partially discharged at t_{int} , proportional to the input luminance; its response will therefore resemble that of a linear pixel. A very bright pixel will be fully discharged before t_{int} , making it impossible to register any further increase in luminance; it has reached the saturation level of its response curve. However, this bright pixel could resume registering photons if a reset voltage pulse V_1 is applied at the time $t_1 < t_{int}$ to replenish the charge of each capacitor. The reset is only partial; the reset voltage V_1 is chosen to be lower than the initial V_0 so as to discriminate between darker pixels that would not be saturated at t_{int} and brighter ones that would. The partial reset does not affect those darker pixels that are still above the reset level V_1 . It only recharges

the brighter pixels that are already discharged below V_1 . For them, the reset overwrites any memory of previous exposure, and they resume registering photons in the shorter exposure time $t_{\text{int}} - t_1$. The corresponding pixel response curve will be piecewise linear, but with a lower incremental gain after the breakpoint that corresponds to the reset at t_1 . The nearer t_1 is to t_{int} , the lower the incremental gain after reset, and the higher the saturation luminance L_{max} becomes. The reset allows a higher luminance range to be “compressed” into the same limited output image signal range; thus, the ratio of the total exposure time t_{int} to the remaining exposure time $t_{\text{int}} - t_1$ is referred to as “compression.” To increase compression further, additional partial resets of successively decreasing voltage can be applied within the remaining time between t_1 and t_{int} , see Fig. 11a,



True Color Night Vision Video Systems in Intelligent Vehicles. Figure 11

(a) Charge-time diagram and (b) response curve for a CMOS WDR pixel with multiple partial resets [5]

thus creating a piecewise linear response curve where the incremental gain is further decreased after each of the breakpoints, see Fig. 11b. Theoretically, the number of additional resets is only limited by the smallest increments of timing and recharge voltage control, but in reality, it is limited by the requirement of detection reliability.

Extending Dynamic Range Reliably

The partial resets of a multiple-slope WDR imager require precise control of the timing and recharge voltage so that it can adapt optimally to the changing dynamic range of a natural scene. Experimental observations and model calculations show how the iSNR criterion can be used to maximize dynamic range without compromising detection of scene detail. The methodology for measuring iSNR has been described in detail [5].

Experimental Observations One important goal of dynamic range extension is minimizing the number of saturated pixels and thus maximizing the capture of highlight detail. With the single-reset system shown in Fig. 10, high compression can be achieved by moving the reset time t_1 as close as possible to the frame end t_{int} . The highest compression is limited only by the increments of reset timing, given by one row time in common multiple-slope imagers and by one pixel time for the most advanced multiple-slope imagers. This can be explained in the example of a VGA imager with 480 rows of 640 pixels. The voltage reset can be placed as close as one row time before t_{int} (i.e., after 479 rows), achieving a maximum compression of $c = 480 / (480 - 479) = 480$. This means that the ratio of exposure times in compressed and linear modes is 1:480 and that the pixel is $480\times$ less sensitive in compressed mode. From exposure time ratios, dynamic range can be estimated to be extendable by 54 dB. In an advanced VGA imager where single pixel times are addressable, the dynamic range extension could be even more impressive. Each row contains 640 pixels. The shortest exposure time would be one pixel time before $t_{\text{int}} = 480 \times 640$ pixels so that a maximum compression of $c = (640 \times 480) / (640 - 639) = 307,200$ could be reached, extending the dynamic range by 110 dB.

Experiments like those shown in Fig. 12 and the corresponding imager response model of section

“Imager Response Model” show that although the practical implementation of high compression with just one partial reset can effectively prevent highlight saturation, it does come at the cost of usefulness, seen as severe loss of midtone image contrast. Figure 12a shows an image taken in linear mode where saturation prevents visibility of detail on the trunk of the car. Progressing from Fig. 12b to d, increased compression decreases the proportion of saturated pixels, making highlight detail more visible on the trunk but at the cost of lowered contrast of midtone detail (lane markings versus road surface). Figure 12d shows the limit of a single-reset system. There are still a fair number of saturated pixels, but the lane markings have become almost indistinguishable from the road surface, which has lost all texture detail. There is a trade-off; the higher the compression, the lower the midtone contrast. At levels of compression still far from theoretical limits, portions of midtone detail almost disappear, and

the image becomes unnatural and less useful. The underlying mechanism of this is explained by the following imager response model.

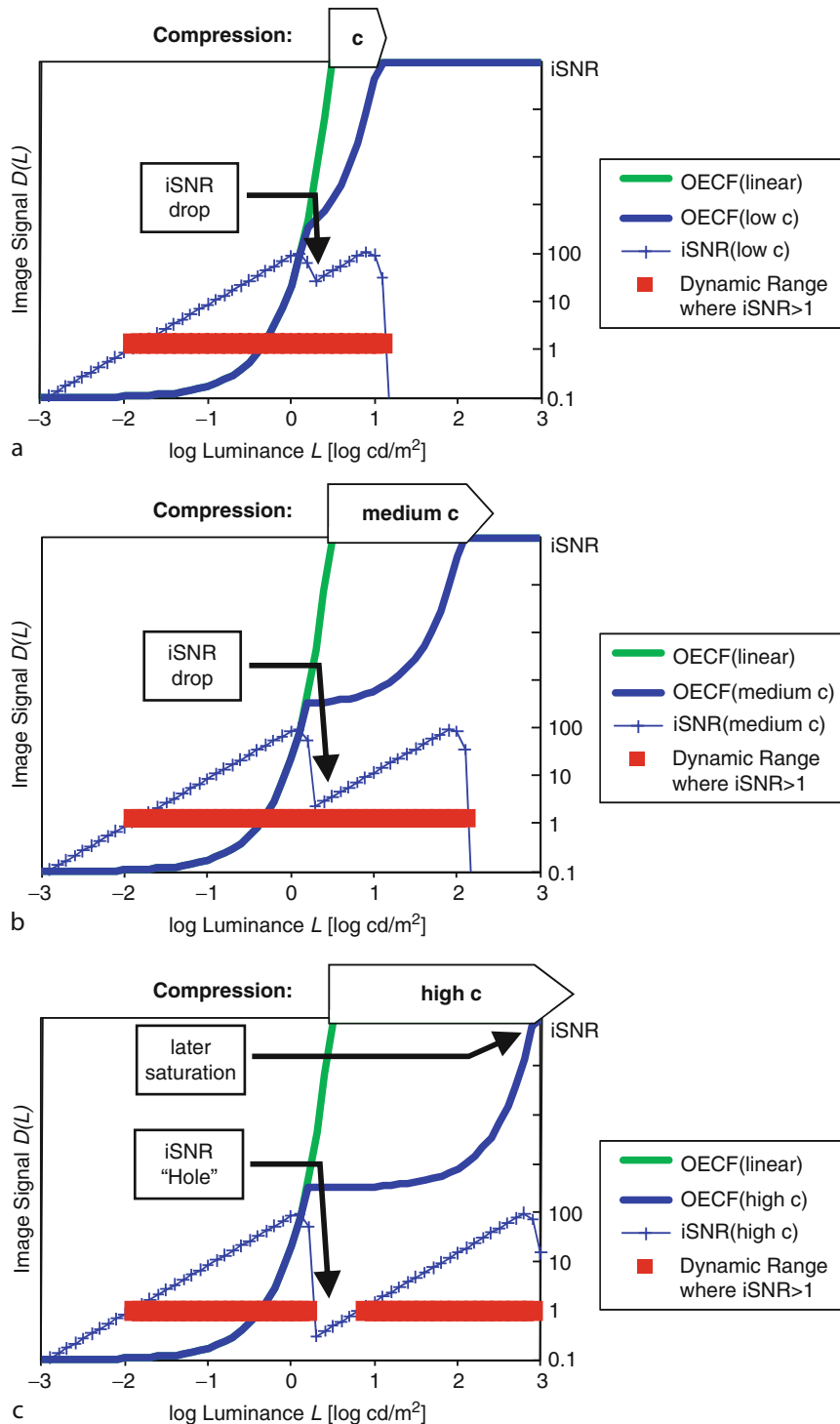
Imager Response Model Response curve and noise models show that the loss of midtone detail as observed in Fig. 12 can be explained by a drop of iSNR that occurs at the reset point. Figure 13 shows modeled curves of OECF and iSNR as functions of log luminance. The model assumes a constant dark noise floor without considering photonic shot noise.

After the partial reset breakpoint, the slope of the response curve (and thus the incremental signal) is reduced so that saturation is delayed to a correspondingly higher luminance level. With increased compression, the drop in iSNR at the breakpoint becomes deeper until it drops below the reliability threshold as shown in Fig. 13c, thus creating an undesirable zone of extinguished local



True Color Night Vision Video Systems in Intelligent Vehicles. Figure 12

Natural scene captured with (a) linear camera and with a single-reset WDR camera at (b) low, (c) medium, and (d) high compression, demonstrating the trade-off between increased dynamic range and decreased midtone contrast [5]



True Color Night Vision Video Systems in Intelligent Vehicles. Figure 13

Log-linear response curves (OECF) and incremental SNR of a single-reset model at (a) low compression, (b) medium, and (c) high compression [5]

contrast, a so-called iSNR hole. The luminance range within and near the “iSNR hole” corresponds to the luminance levels of those details such as road texture and lane markings that became almost indistinguishable in Fig. 12d.

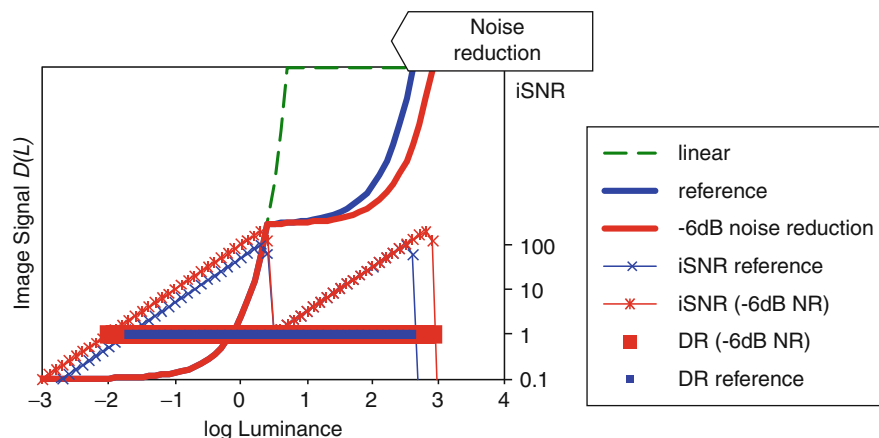
Along with saturated pixels or those below the noise floor, “iSNR holes” make the detection of scene detail impossible. In linear imagers, the zones of iSNR failure occur only at very dark pixels (below noise floor) and very bright pixels (above saturation level). However, in multiple-slope WDR imagers, zones of iSNR failure can also occur at luminance levels corresponding to reset times. These levels cannot be predicted without exact knowledge of the response curve for a given set of reset times. If the resets are automatically adapted to the intra-scene dynamic range, then the luminance levels corresponding to the reset times would depend on scene content and thus be variable, making iSNR holes potentially interruptive to the visibility and automatic detection of scene detail. It follows that resets have to be controlled to prevent iSNR holes occurring in the first place. Dynamic range can only be extended as long as the incremental signal iS can be kept above the noise floor at every luminance level. Thus, iSNR is not just a criterion for determining the upper and lower dynamic range limits but becomes critical when controlling the reset times within the dynamic range. Adaptive WDR extension requires separate control loops for integration time and compression that both adapt the imager’s response curve to include the

darkest and brightest scene details. The control algorithms for timing the multiple partial resets in adaptive WDR extension must be designed specifically to avoid iSNR holes.

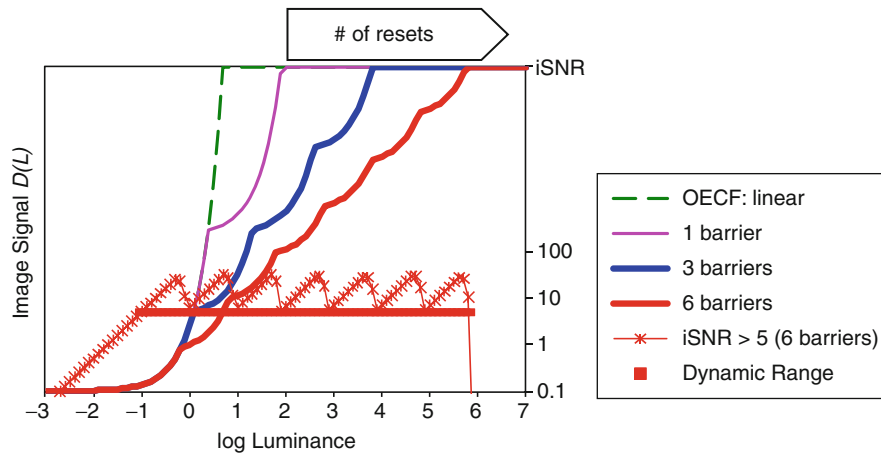
Reliably Extending Dynamic Range The example of a multiple-slope WDR camera demonstrates how the iSNR criterion can be used to extend dynamic range without sacrificing detection reliability.

If the number of resets is limited to one or two, then noise reduction is the most effective way of increasing dynamic range. Figure 14 compares two single-reset model imagers, one where altered pixel design (i.e., larger size, correlated double sampling, etc.) has cut noise in half (−6 dB). This leads to a twofold increase in dynamic range. Firstly, the 6 dB lower noise floor leads to a 6 dB lower noise-equivalent luminance signal $L_{\min}$. In addition, the lower noise floor allows – without creating an iSNR hole – the compression to be increased by a factor of 2, halving the slope after reset and doubling the saturation luminance $L_{\max}$, thus increasing the dynamic range by another 6 dB, a total of 12 dB. Assuming a minimum iSNR of 1, dynamic range is increased from 88 to 100 dB or twice that of a linear imager of 50 dB.

Where pixel-level noise reduction is not effective (small pixels, low fill factor, high operating temperatures, inexpensive pixel design), then increasing the number of partial resets is a better strategy. The drop in iSNR at a reset point depends on the ratio of



True Color Night Vision Video Systems in Intelligent Vehicles. Figure 14
Lower pixel-level noise increases dynamic range in a single-reset imager [5]



True Color Night Vision Video Systems in Intelligent Vehicles. Figure 15

More resets increase dynamic range while allowing a higher minimum iSNR threshold of 5 [5]

incremental gains after and before the reset, $k = g/g_0$. The more $g(L)$ is reduced by a reset, the deeper iSNR will drop. After each partial reset n , the sensitivity decreases geometrically by the factor $1/k^n$. When deploying multiple partial resets, their timing and charge (voltage) can be chosen so that the linear output signal range is equally distributed between the resets [5]. This ensures that iSNR at each reset point drops to the same minimum level (see Fig. 15). Having more resets has the added advantages: the decrease in incremental gain can be approached more gradually, and the iSNR threshold can be raised from one to a more practical level. Figure 15 shows a dynamic range model for an increasing number of resets. Assuming an increased iSNR threshold of 5, the linear imager will have a dynamic range of only 36 dB. Compression with a single reset will extend the dynamic range to 62 dB; two resets will reach 98 dB, and six resets 138 dB.

Conclusion

Although the concept of iSNR-based reliability has been modeled and tested only for multiple-slope CMOS imagers, it can be deduced that the concept applies to any other WDR extension technology that involves sudden changes in the incremental gain of the pixel response curve. In contrast, sensors with log-arithmetic response are in no danger of “iSNR holes” as their incremental gain decreases continuously during exposure time.

For the example of multiple-slope CMOS imagers, it has been shown that iSNR is a critical criterion governing the number, timing, and recharge voltage of partial resets. Dynamic range can be increased by reducing noise at the pixel level, by increasing the number of partial resets, and by distributing them equally to maintain a minimum iSNR necessary for reliable detail detection throughout the entire dynamic range. As technologies for extending the dynamic range evolve, new solutions will be added to the repertoire and tested for their suitability for intelligent vehicle technology. However, the methodology described here will provide a guideline for optimizing the pixel response for maximum dynamic range at the highest possible iSNR.

Extending Spectral Sensitivity into NIR

In order to capture dark scene detail illuminated by ambient light, the camera has to be very sensitive with high SNR. Exposure time is limited by the video frame rate and the need to limit motion blur. Sensitivity can be increased by either lowering noise (increasing SNR) and/or extending the range of spectral energy sensed.

IR Sensing

The spectral sensitivity of silicon-based image sensors extends beyond the long-wavelength edge of the visible spectrum (730 nm) into the infrared (IR). As longer

wavelengths penetrate further into silicon, CMOS imagers can be made more sensitive to IR by collecting photo charge carriers from greater depth and by increasing substrate thickness. The IR cut-off in CMOS imagers can thus be shifted as high as 1,100 nm. This spectral range coincides with the shorter wavelengths of the IR-A or near-infrared (NIR) range (700–1,400 nm). Imagers for longer wavelengths beyond the NIR rely on semiconductors with a narrower band gap, for example InGaAs (950–2,600 nm). Such far-infrared (FIR) cameras can sense the black body radiation that surfaces emit as a function of their temperature, and obtain a thermal image of the environment. They do not require any other source of radiation and can operate in complete darkness. Compared to NIR systems, FIR night vision systems are more costly, cannot be placed behind glass, and deliver unnatural images that are difficult to interpret [27, 28].

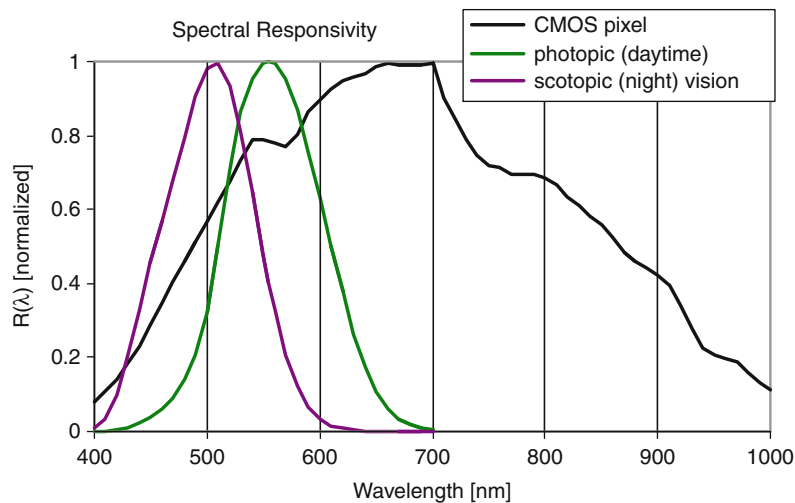
Passive and Active NIR Vision

The detection of objects by a NIR-sensitive camera requires them to be irradiated by a NIR source so that the camera can detect radiance image from reflected NIR. Passive sensing relies on ambient NIR. Ambient NIR at night comes mostly from artificial incandescent light sources such as vehicle headlights; vapor street

lighting and natural light sources (moon, stars) emit comparatively little NIR. Daylight especially direct sunlight contains extremely high background NIR that is strongly reflected by vegetation; cameras may have to operate in WDR mode to avoid signal saturation. Active sensing requires sources that are mounted to the same vehicle as the NIR camera. Conventional metal halide (tungsten) headlamps emit more than 60% of their spectral energy in the NIR. Dedicated NIR-only headlamps, either tungsten lights with a VIS-cut filter or NIR LEDs at 850 nm, have the advantage that they are invisible to humans. As they cannot blind oncoming drivers, they can be operated on high-beam, thus extending the range of vision beyond that of dipped headlights [29]. One of the first active systems implemented was the Mercedes-Benz Night View Assist (2005).

Comparing the spectral sensitivity of a monochrome CMOS image sensor with the photopic and scotopic luminous sensitivity functions shows the range of spectral energy that the image sensor can utilize. For example, when the scene is illuminated by metal halide headlamps whose emission spectra peak in the NIR at about 900 nm, sensitivity stretches to the long-wavelength cut-off of the image sensor (Fig. 16).

The example in Fig. 17 shows an urban traffic scene taken without and with ambient NIR, showing the main



True Color Night Vision Video Systems in Intelligent Vehicles. Figure 16

Comparison of photopic and scotopic luminous sensitivity with the spectral sensitivity of a monochrome automotive CMOS image sensor [42]



True Color Night Vision Video Systems in Intelligent Vehicles. Figure 17
Monochrome image of urban traffic scene (a) without NIR and (b) with NIR

benefit of extended NIR sensitivity: better visibility of the roadway where it is illuminated by dipped headlights.

The range of visibility can be extended even further, without blinding oncoming drivers, if dedicated NIR-only high-beam headlights are used. These will, of course, “blind” any NIR-sensitive night vision cameras of other vehicles, unless they are capable of performing in a WDR environment.

Adding Color Differentiation to VIS + NIR-Sensitive Imager

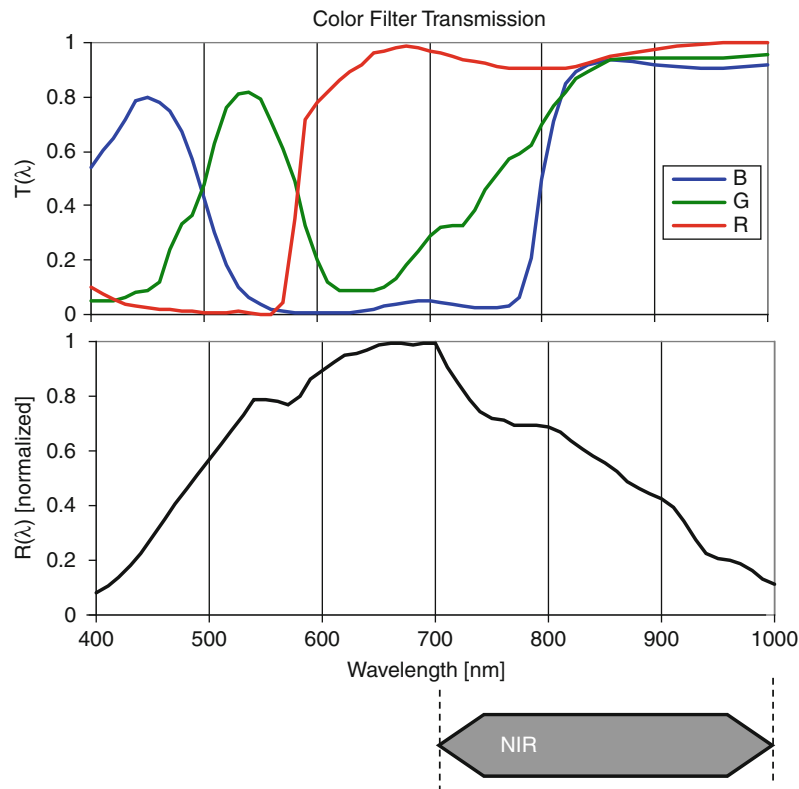
Using NIR to increase the sensitivity of color imagers requires separation of the NIR and RGB color signals because the color filter materials used in standard CMOS and CCD technology are highly transparent to NIR so that each color pixel also detects NIR (Fig. 18).

Mixing color and NIR signals can result in extreme color de-saturation if the illumination contains high

amounts of IR. This is the case when incandescent lamps are used with color temperatures typically about 3,200 K, where spectral power distribution peaks at about 800 nm (Fig. 7). Here, the signal contributions to RGB from the NIR are even higher than those from the actual colors so that all colors are rendered as off-gray, for example shades of pink (Fig. 19). The example demonstrates that the differentiation of color requires the NIR signal to be separated from the color signals.

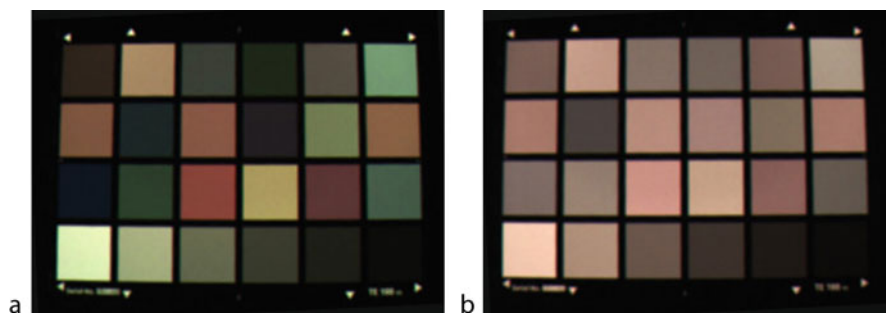
Increasing the Spectral Sensitivity of Color Imagers

Color image sensors simply suppress the NIR information by an IR-cut filter. The sensitivity of such imagers can be improved by adding filterless “white” pixels to the RGB filter pattern [30, 31] (Fig. 20). Compared to the RGB color pixels, these “white” pixels have about three times higher spectral sensitivity.



True Color Night Vision Video Systems in Intelligent Vehicles. Figure 18

VIS-NIR spectral transmission curves of typical RGB color filters used in CMOS imagers [35]



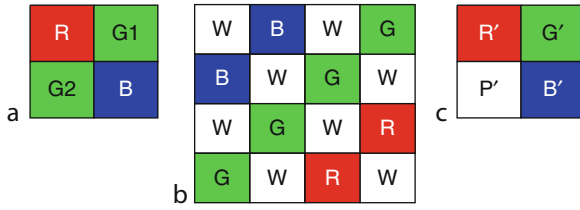
True Color Night Vision Video Systems in Intelligent Vehicles. Figure 19

Raw images of a Macbeth color checker chart illuminated by incandescent light (3,200 K) taken with a CMOS color camera (a) with IR-cut filter and (b) without IR-cut filter [35]

As scene illuminance decreases, information from color pixels is increasingly mixed with luminance information from the W-pixels by pixel binning. This requires high pixel counts plus special (nonstandard) filter patterns, and the NIR remains unutilized for night vision.

Extending the Spectral Sensitivity of Color Image Sensors into the NIR

Extending the light sensitivity of color imager sensors into the NIR requires a separation of the NIR and the color signals. This could be achieved by using a beam



True Color Night Vision Video Systems in Intelligent Vehicles. Figure 20

CFA layouts: (a) standard Bayer, (b) RGBW, and (c) RGBP [35]

splitter and two image sensors, one monochrome for NIR, and one color for VIS. However, cost and form factor do not recommend this solution for automotive vision applications.

Single-imager solutions are based on the addition of a fourth pixel type to the RGB-pixels (Fig. 20c). This fourth pixel, which, for example, replaces the second G-pixel in the Bayer pattern, is either NIR-sensitive or “panchromatic,” sensitive to both visible and NIR light. A variety of methods have been proposed to separate NIR signals from color signals:

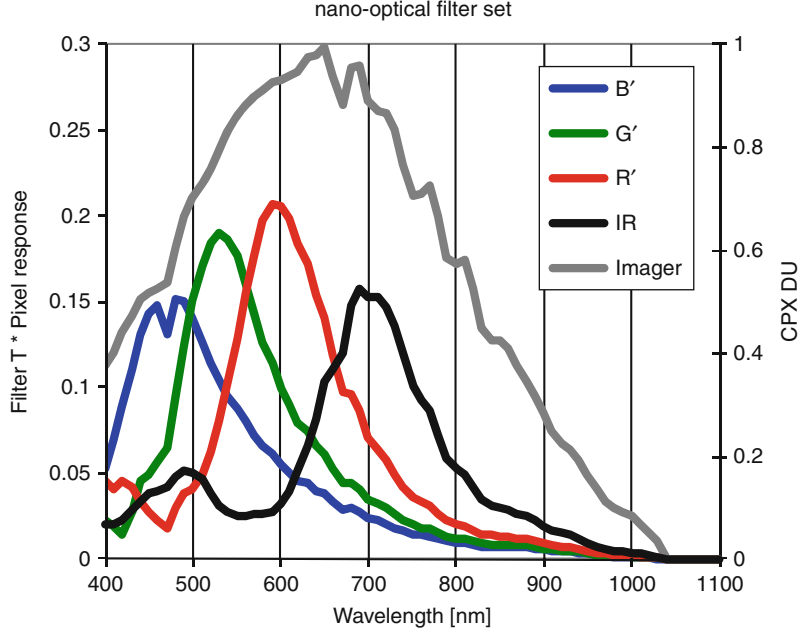
- Patterned optical NIR-cut filters can be added to mosaic filter arrays, for example RGBW or CymG (cyan-yellow-magenta-green) [32], to protect all or a portion of color pixels from the effects of NIR radiation. The filters can be of the absorptive or interference type. Interference filters are preferred for their sharper absorption edge toward the NIR, but they do require patterned deposition of more than ten dielectric layers [36], which is both costly compared to dye-based filters and uses dielectric materials that are incompatible with standard Si processes. The effectiveness of such filters is limited by optical crosstalk of NIR into the color pixels after passing the NIR cut filter, especially at low incident angles.
- Nano-optical filters, currently still in the research stage, can be designed as short-pass edge filters or band-pass filters. They can either be integrated as metal layers into the pixel stack [37] or deposited on top of the imager [38]. Although nanostructured metal layers within the pixel stack have to follow the Si design rules for structure size (0.15, 0.13, 0.11, 0.09 μm) which limits the degrees of freedom for spectral tuning, they are practically

free from optical crosstalk. Filters on top of the imager can be created by nanoimprint technologies which allow the creation of an array of plasmonic band-pass filters. However, the peak transmission of band-pass filters is relatively low (<0.4). Peak filter transmission decreases, and halfwidth increases with wavelength, making nano-optical filters less effective in the NIR (Fig. 21) [38].

- Virtual NIR filters separate NIR and color signals by image signal processing. One possible implementation adds filterless pixels to the RGB filter pattern, but in contrast to RGBW, no optical NIR-cut filter is used [33–35]. The filterless pixel has two functions: Firstly, it has a “panchromatic” spectral response that includes the NIR and is equivalent to that of a monochrome imager. The panchromatic P'-pixel is more than three times as sensitive as a color imager, capable of delivering a much brighter luminance image. Secondly, the P'-pixel delivers the reference signal necessary to estimate the NIR signal contribution to the $R'G'B'$ pixel signals. The raw color pixel responses are $R' = R + I$, $G' = G + I$, and $B' = B + I$ (Fig. 22). The panchromatic pixel response $P' = R + G + B + I$ provides the reference signal for estimating the NIR portion in the incident light, which is then used in the virtual NIR filter matrix to remove as much as possible of the NIR contributions to the $R'G'B'$ signals. The response of the P'-pixel can also be used to simply retrieve a panchromatic VIS–NIR luminance channel $L_{\text{VIS-NIR}}$.

An example of raw spectral pixel responses ($R'(\lambda)$, $G'(\lambda)$, $B'(\lambda)$, $P'(\lambda)$) shows the high NIR signal contribution to the color signals as well as high amounts of crosstalk between the four pixel types. Since both deteriorate colors, they have to be minimized by determining the virtual NIR filter coefficients, for which two calibration methods are commonly used, global and spectral.

- The global method relies on a set of known reference colors, for example those of a Macbeth color checker chart, that have to be reproduced as faithfully as possible. The color match can be done in a standardized color space, for example CIE 1931 XYZ, and the matrix M_G optimized so that the



True Color Night Vision Video Systems in Intelligent Vehicles. Figure 21

Simulated response curves for CMOS imager with plasmonic nano-optical filters [38]

transformed colors match the reference colors $[XYZ]$ as closely as possible,

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \approx M_G \cdot \begin{bmatrix} R' \\ G' \\ B' \\ P' \end{bmatrix}. \quad (5)$$

Global calibration works well for colors within and close to the reference set and under illumination conditions similar to those used in the calibration, but it is poorly suited for color night vision where the most relevant colors are not reflective but rather emissive colors whose brightness might lie far outside the set, and whose illuminant, and thus, relative NIR content varies widely.

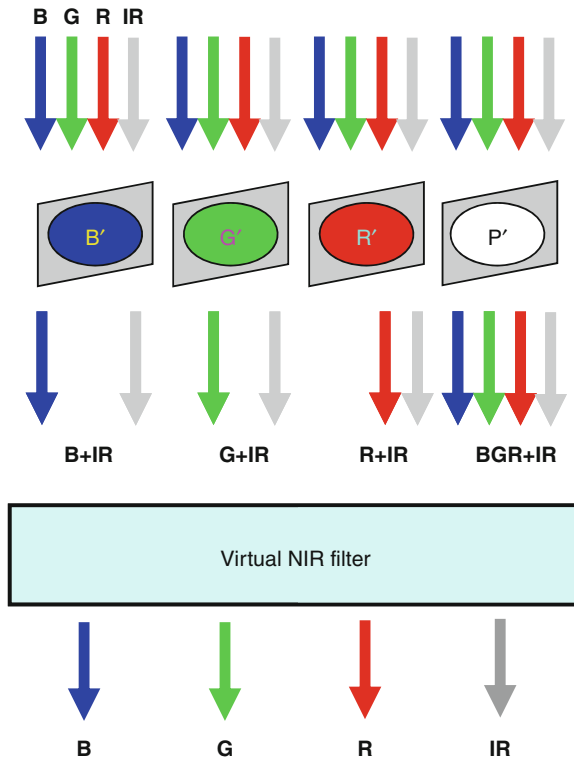
- The spectral calibration method optimizes the matrix coefficients β_{mn} so that the filtered spectral response functions $(R(\lambda), G(\lambda), B(\lambda), I(\lambda))$ fulfill application-specific conditions.

$$\begin{bmatrix} R(\lambda) \\ G(\lambda) \\ B(\lambda) \\ I(\lambda) \end{bmatrix} = \begin{bmatrix} \beta_{11} & -\beta_{12} & -\beta_{13} & \beta_{14} \\ -\beta_{21} & \beta_{22} & -\beta_{23} & \beta_{24} \\ -\beta_{31} & -\beta_{32} & \beta_{33} & \beta_{34} \\ \beta_{41} & \beta_{42} & \beta_{43} & -\beta_{44} \end{bmatrix} \cdot \begin{bmatrix} R'(\lambda) \\ G'(\lambda) \\ B'(\lambda) \\ P'(\lambda) \end{bmatrix} \quad (6)$$

For example, in a machine-vision application, the coefficients are optimized to remove from the transformed spectral responses $(R(\lambda), G(\lambda), B(\lambda), I(\lambda))$ as much of the NIR and color crosstalk signals as possible. Instead of retrieving an IR channel, the P' - response can be mapped into a panchromatic luminance channel $L_{\text{VIS-NIR}}$. In human vision applications, a colorimetric calibration is better suited for achieving natural-looking colors. Colorimetric calibration as shown in Fig. 23c tries to optimize the matrix M_S so that the filtered spectral response curves of the imager approximate the spectral color matching functions $(b(\lambda), g(\lambda), r(\lambda))$ of the human visual system as specified in ISO 17321 [39]:

$$\begin{bmatrix} r(\lambda) \\ g(\lambda) \\ b(\lambda) \end{bmatrix} \approx M_S \cdot \begin{bmatrix} R'(\lambda) \\ G'(\lambda) \\ B'(\lambda) \\ P'(\lambda) \end{bmatrix}. \quad (7)$$

Spectral calibrations have the advantage of not relying on a particular set of reference colors, and also, they do not rely on a specific illuminant. The virtual NIR filter with colorimetric spectral calibration reduces the NIR crosstalk into color signals even under



True Color Night Vision Video Systems in Intelligent Vehicles. Figure 22

Light, pixel responses of the $R'G'B'P'$ imager and image signals after virtual NIR filtering [35]

illumination conditions with high amounts of NIR, for example incandescent light at 3,200 K color temperature. The example in Fig. 24 shows images before (a), (b), and after virtual NIR filtering (c) with increasing amounts of NIR. Increasing NIR reduces color saturation of the raw $R'G'B'P'$ color channels, and increases brightness of the panchromatic P' -channel. The colors reconstructed by the virtual NIR filters become darker with increasing levels of NIR because of decreasing color-to-NIR signal ratios in the raw images.

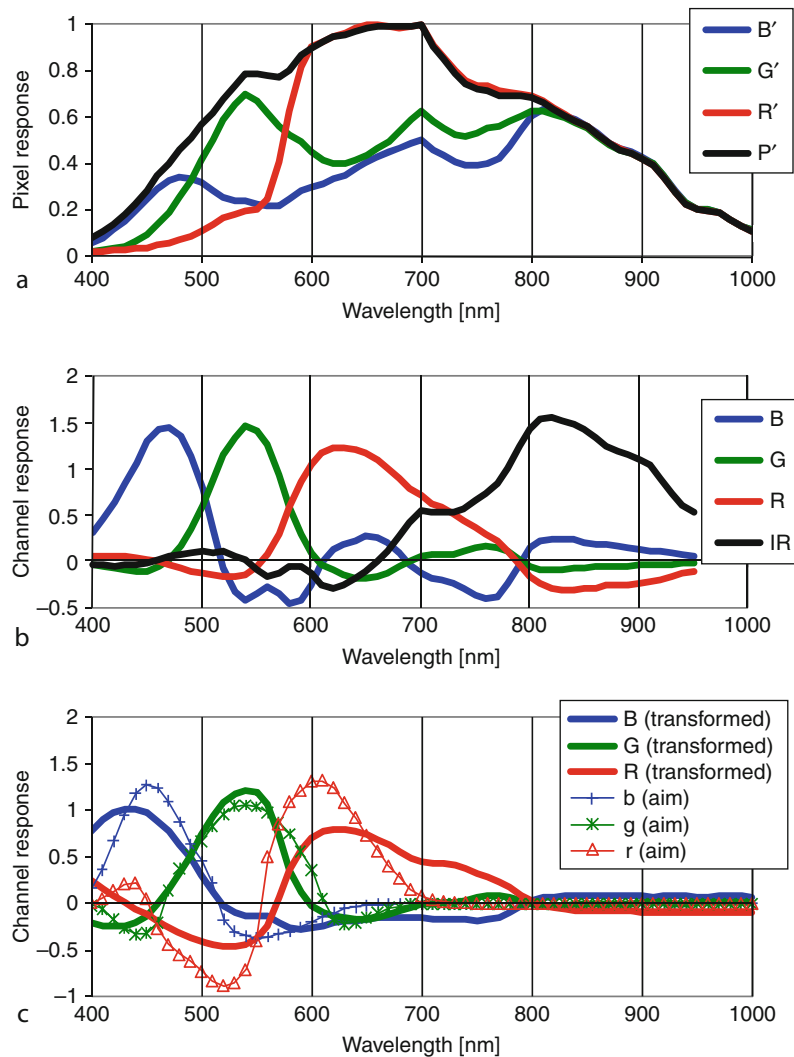
However, residual spectral leakage limits the performance of the virtual NIR filter, especially in WDR mode where initial signal ratios are reduced. Careful management of NIR can further reduce the effects of NIR crosstalk and has to include pixel and microlens design to minimize optical and electrical crosstalk, high SNR to counter the noise-amplifying effect of high filter coefficients, and NIR-only headlights to

minimize the proportion of NIR radiation with the highest leakage into the color signals.

Extending the Dynamic Range of Color Imagers with NIR Sensitivity

Signal saturation in color imagers leads to loss in color information; as the signals in the R, G, and B channels approach saturation levels, colorfulness decreases. In a RGB color imager, signal clipping in one or more color channels leads to a whitening of colors; clipping in a four-channel system such as $R'G'B'P'$ can also lead to false colors. WDR extension is therefore essential to capture color detail carried by bright lights. However, corrections have to be made for the effects of WDR extension on the color signals. In the example of multiple exposures with varying exposure time, each exposure maps a different range of input radiance signals into the same range of output image signals. Within the sequence of images to be combined into a single WDR image, the same output signal may be produced by different combinations of input signals so that the relationship between input and output signals is not unique. Therefore, the combination of exposures into the WDR image has to be preceded by linearization, which uses the camera response function in each color channel to create a unique map between the digital signal levels in the WDR image and the input radiance levels. The response function can be derived by techniques such as those proposed by Debevec and Malik [40], or Mitsunaga and Nayar [41].

Direct WDR methods that extend the dynamic range in a single exposure have the great advantage of providing a unique map between input radiance and output image signal levels (Fig. 9). However, if the response function is nonlinear, for example piecewise linear as shown in Fig. 11, the signal ratios between the color channels will decrease with decreasing slope of the response curve. Increasing compression will thus decrease the input signal ratios between the raw $R'G'B'P'$ channels, thus decreasing chromaticity and iSNR of the output colors after virtual NIR filtering. To recover the color components and remove the NIR component from the highly compressed signals of bright colored lights, the raw $R'G'B'P'$ pixel signals have to be linearized with the inverse response curve of the imager so that the image signals can be referred



True Color Night Vision Video Systems in Intelligent Vehicles. Figure 23

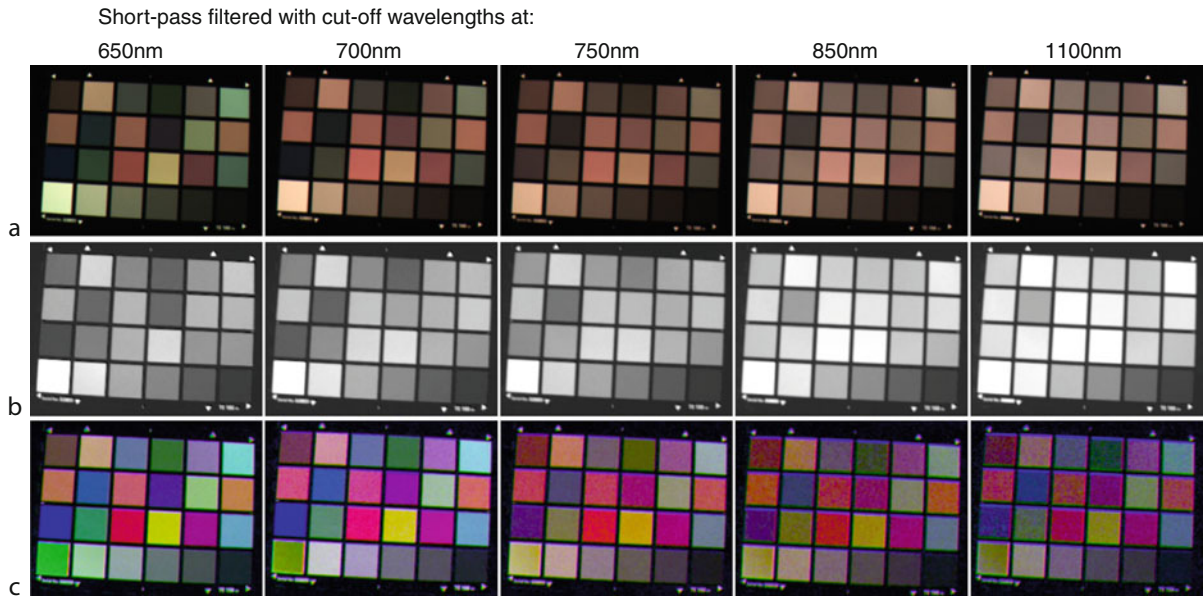
Spectral pixel responses of (a) the R', G', B', and P' pixels and spectral channels after (b) separating NIR from the RGB color channels and (c) colorimetric calibration [42]

back to the corresponding scene radiances (inverse gamma correction). The effect of linearization is shown in Fig. 25 where it restores the colors in the highly compressed images of bright colored lights below the Macbeth color checker chart [42].

Visual Representation of the VIS + NIR Color Information

Section “Conclusion: Producing Natural Images Under Mesopic Conditions” outlined the requirements for

rendering a natural-looking night vision image. In the example of a WDR imager with panchromatic and color pixels, NIR sensitivity increases the range of visibility, and the panchromatic image information can be displayed in the luminance channel. The virtual NIR-cut filter can recover traffic-relevant color information from backlit or retroreflective signage, traffic lights, and taillights. Virtual NIR filtering shows reasonably good performance in reconstructing colors even under illumination conditions with high amounts of NIR, for example incandescent light at a color temperature of



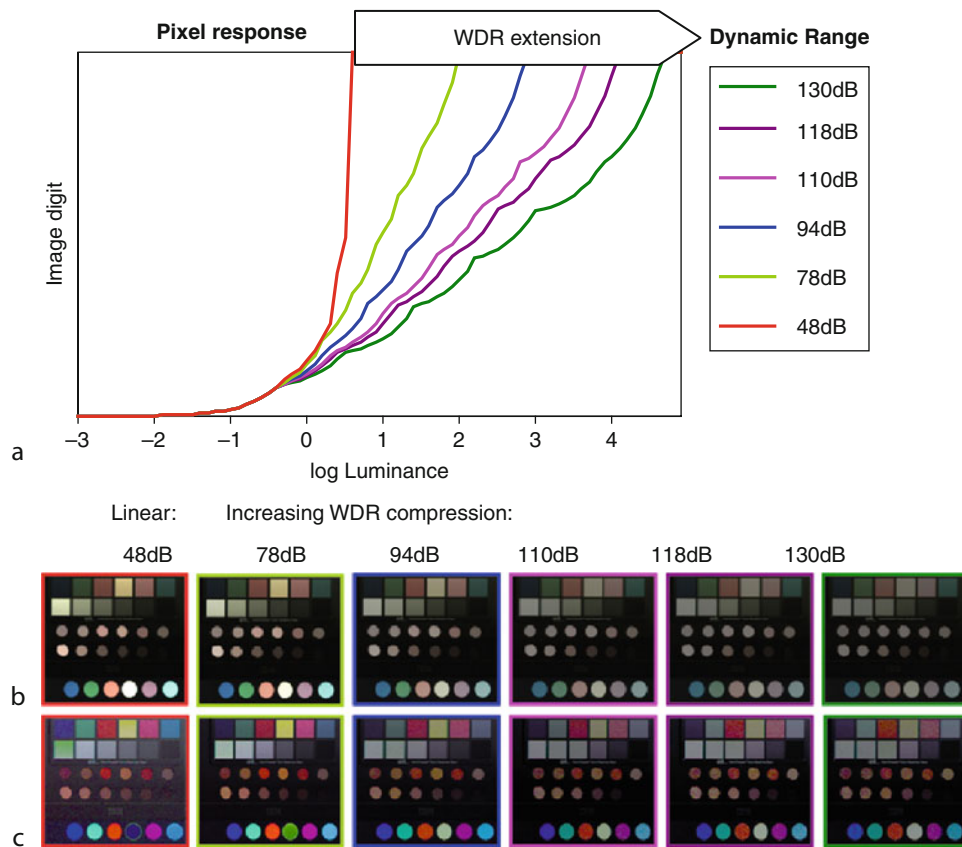
True Color Night Vision Video Systems in Intelligent Vehicles. Figure 24

Macbeth color checker chart illuminated by incandescent light (3,200 K) and imaged with short-pass filters allowing increasing amounts of NIR: (a) raw R'G'B' image, (b) raw P' image, and (c) RGB image after virtual NIR filtering [42]. The false colors in the brightest patch are caused by clipping of the P'-signal

3,200 K having 20% signal in the visual and 80% in the NIR range (Fig. 24). Since the filter array relies on standard RGB filter materials and fabrication processes, it can be produced at comparably low costs as conventional RGB imagers, without the extra cost of a patterned optical NIR-cut filter.

The VIS and NIR image information has to be represented in a way that the human observer viewing video on a dashboard screen can easily interpret. The VIS + NIR color camera with virtual NIR filter can provide multispectral image information, for example, in four channels, RGB and NIR. Multichannel image data could simply be mapped into a false-color representation, but this creates unnatural images that do not agree well with visual expectations and can even be confusing rather than useful. Natural image data representation that agrees with human expectations can be achieved by following certain design guidelines [43]: (1) orthogonal encoding to represent information with as little redundancy as possible; (2) anthropometric optimality to maximize usefulness to the human observer; (3) ecological invariance so ensure similarity of image interpretation over a variety of environmental parameters such as time

of day, illuminants, and weather conditions; (4) computational simplicity to allow video implementation. For example, the achromatic channel of the HVS carries over 95% of the scene information and is much more sensitive to spatial information than the opponent color channels which process color information. Therefore, true-color image fusion is best suited for this application because it uses the extra information from the NIR to enhance the spatial and luminance information in the achromatic channel while mapping the color information from the visible range into the chrominance channels [32, 44, 45]. First, the RGB image obtained by virtual NIR filtering (6) must be transformed into a luminance–chrominance color space; then, its luminance component L_{VIS} is replaced by the luminance $L_{\text{VIS+NIR}}$ estimated from the P'-pixels. Additional processing of $L_{\text{VIS+NIR}}$ ensures that the fusion increases the brightness of the dark scene background but not so much that of colored lights, thus preventing the loss of color saturation in colored lights after fusion. For example, tone mapping using global, local, or Retinex algorithms significantly improves the visibility of grayscale and color detail in WDR images [46].



True Color Night Vision Video Systems in Intelligent Vehicles. Figure 25

WDR color scene consisting of dark reflective colors (*top*), bright transmissive colors with (*center*) and without NIR (*bottom*), taken with (a) imager response of increased compression and displayed (b) without and (c) with linearization [42]

The night scene in Fig. 26, illuminated only by the car's headlights, shows how image fusion effectively extends the field of vision, increases visibility of detail, and retains color in the retroreflective sign.

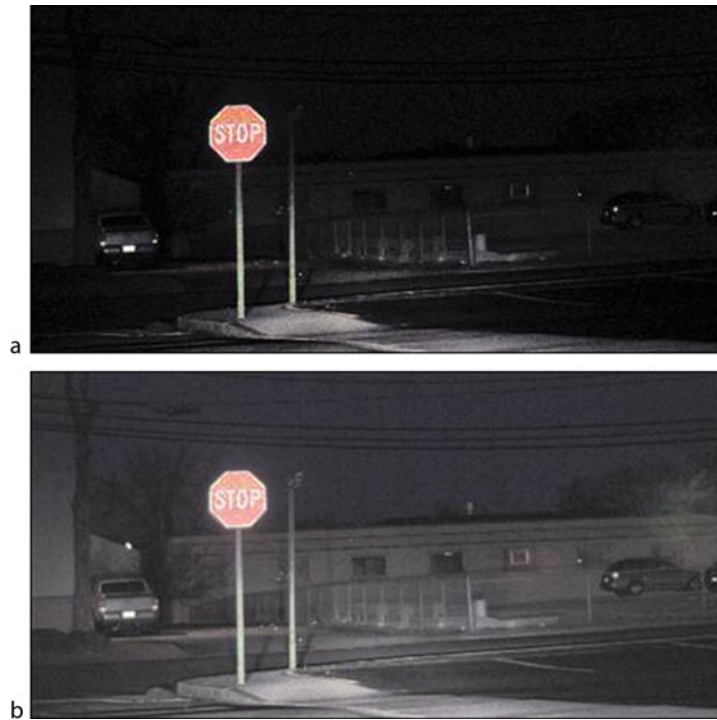
A previously described WDR scene simulator [5] composed of a digital display and a projector has been used for creating test scenes with controlled luminance levels, covering a wide dynamic range of 100 dB. Individual scene elements such as the road surface, lane markings, traffic signs, pedestrians, and headlights shown in Fig. 27 were calibrated so that their absolute luminance levels were within the ranges of typical traffic scenes as shown in Fig. 3. The example demonstrates that even if little or no IR is present in the scene, image fusion will still increase the brightness of dark scene detail because the filterless P'-pixels are more sensitive than any of the color pixels. Fusion utilizes the brighter luminance

image from the P'-pixels, making the pedestrians in the dark background much apparent. The retroreflective sign and traffic lights are also brightened, but only to the extent that color saturation is not adversely affected.

In real-world night vision scenes, the panchromatic pixel can utilize the large NIR component of incandescent (tungsten) headlights that is invisible to the driver (Fig. 7). Furthermore, high-beam NIR-only headlights can be used to increase visibility beyond the distance range of headlights at low beam.

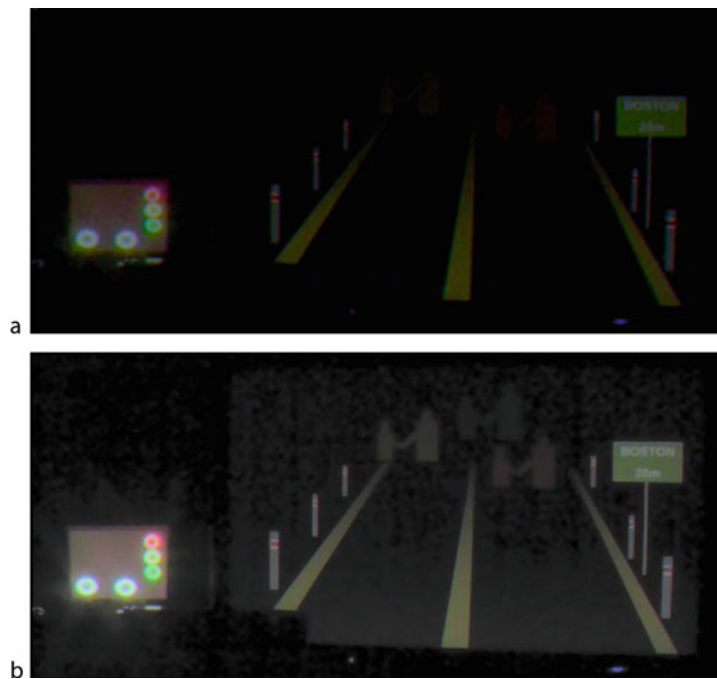
Example: Render Traffic-Relevant Color Information in Natural Images

Figure 28 demonstrates the image rendering of a VIS-NIR true-color WDR night vision camera for a typical urban night traffic scene.



True Color Night Vision Video Systems in Intelligent Vehicles. Figure 26

Night scene with tungsten headlights as only light source, taken without image fusion (a) and with image fusion (b)



True Color Night Vision Video Systems in Intelligent Vehicles. Figure 27

Simulated 100 dB WDR traffic color scene (a) without and (b) with image fusion [35]



True Color Night Vision Video Systems in Intelligent Vehicles. Figure 28

Urban night traffic scene: (a) panchromatic P' image, (b) RGB color image after virtual NIR filtering, and (c) final image after image fusion [8]

The monochrome image in Fig. 28a shows that the wide spectral sensitivity of the panchromatic P' -pixels delivers a sufficiently bright monochrome image even at a frame rate of 30 fps. The color image in Fig. 28b is comparatively dark, consisting mainly of the “unrelated colors” of traffic lights and

vehicle taillights against an almost black background. The spectral energy in the visible band is small compared to the total energy within the imager’s sensitivity range so that the luminance component of the filtered colors is much lower than the luminance from the P' -pixels. WDR compression and

subsequent linearization ensure that the bright high-light colors are preserved. The fused image in Fig. 28c then combines the bright monochrome background with the unrelated colors, leading to a more natural rendering of the scene that contains all traffic-relevant color information from traffic lights and car signal lights.

Conclusion

True-color night vision systems that utilize the NIR sensitivity and WDR capability of CMOS image sensors can aid the driver in dangerous situations where human vision is inadequate or challenged (Table 2). NIR sensitivity extends the range of vision when the safe braking distance exceeds the range of headlights. Cameras with adaptive WDR compression provide natural images of scenes where the intra-scene dynamic range exceeds that of the human eye, for instance when blinded by headlights of oncoming vehicles. Such cameras are able to adapt to sudden changes in scene brightness much faster than the HVS so that image information is available while the eye is still in the process of adapting, for example to the “black hole” after bright headlights of a passing car have left the field of vision.

True Color Night Vision Video Systems in Intelligent Vehicles. Table 2 Advantages of camera-based night vision systems

	Human visual system	True-color VIS–NIR camera with WDR	Advantages of camera
NIR sensitivity	No	Up to 1,100 nm	Extended range of vision
Adaptation to brightness and dynamic range	Transient 1–2 s	3–6 frames (0.2–0.1 s)	Image information available at WDR and during sudden changes in brightness
Color vision	Mesopic shift	Color constancy	Consistent color information

Future Directions

Vision-based DAS are one of a wide range of intelligent vehicle technologies that can be implemented either as stand-alone or cooperative systems. Some are already in use (antilock braking system, electronic stability control) while others are still in the research stage, under development, or in the process of being introduced into the market [47, 48].

Vision-based driver assistance systems aim to enhance the visual perception of the road ahead in situations where the human visual system is challenged or even fails to perceive dangerous situations. The scene information captured by the camera is processed into a natural-looking image and displayed in the driver's field of view. Since more than 90% of driving decisions and actions are based on vision, the fastest and most efficient human–machine interface is a visual display. Their combination of relative simplicity with high capacity and the speed of the visual information channel recommended display-based vision systems to early proponents of intelligent vehicle technology. Active NIR-sensitive night vision systems such as the Toyota Night View (2002) and Mercedes-Benz Night View Assist (2005) were pioneered in top-of-the-range cars. Despite the relative simplicity of camera-to-display night vision systems versus machine-vision-based automatic driver assistance systems (such as lane keeping, obstacle detection, etc.), the technical challenges and reliability requirements are so high that their penetration into the market is still low. The night vision system must deliver useful natural images under an extremely wide range of lighting conditions while experiencing changing proportions of color and NIR signals. Furthermore, they have to perform under adverse weather conditions that severely tax human vision, such as rain, drizzle, fog, snow, and salt spray. The performance and reliability of future night vision systems might be improved in many ways, perhaps including:

- Improved NIR management by integrating adaptive NIR-only headlights into the night vision system [49]
- Increased SNR of the image sensor to increase dark sensitivity and counter the increased visibility of color noise after virtual NIR filtering

- Reduced NIR crosstalk into the color signal through CFA and color pixel design, for example, nano-optical color filters [37, 38]
- Reduced NIR leakage in virtual NIR filtering by using algorithms that further reduce NIR leakage, for example, neural networks [50]
- Improved WDR color processing, for example, Retinex-based tone mapping [46]

In addition, head-up displays such as windscreen projection will be able to communicate visual information to the driver without shifting attention away from the road as necessitated by a head-down display outside the field of vision, for example, the dashboard displays of early night vision systems [28].

The main shortcoming of any display-based vision system is that it only improves traffic safety if the driver pays attention to the display at the critical moment. For example, it is of no use if the driver does not check the display as an obstacle beyond the field of vision is displayed by the NIR-sensitive system before the driver can see it through the windshield. Recent developments in intelligent vehicle technology work toward addressing this shortcoming and move beyond the early display-based vision systems. The two main strategies are intelligent stand-alone systems to enhance vehicle safety, and cooperative systems to improve road traffic safety [48].

- Stand-alone, in-vehicle systems use sensors to detect possible dangers and try to avoid or mitigate accidents by following a three tier strategy: (1) inform the driver as early as possible, (2) issue warnings if there is no driver reaction to the information, and (3) actively assist or ultimately intervene in the vehicle control [48]. In-vehicle systems will rely on multiple vision sensors and combine different sensor technologies such as radar and LIDAR (light detection and ranging) alongside vision sensors to improve overall detection reliability. Display-based vision systems can be expected to remain an important integral part of on-board safety systems because a visual display more immediately informs the driver of the nature of the danger than an acoustic warning would.
- Cooperative systems increase the capacity for gathering information by using communications from infrastructure or other sensor-equipped vehicles to make an individual vehicle less dependent on its

own sensors and local environmental conditions. For example, if a car drives into a bank of fog, it can be guided by infrastructure and information from other vehicles outside the danger zone.

Forward-looking VIS–NIR cameras with color discrimination and WDR will continue to produce useful image information for both human and machine vision systems. Even if camera-based night vision systems with visual human–machine interfaces turn out to be only a transition stage toward future intelligent vehicle solutions, the technical experience gleaned by the implementation and application of pioneering vision systems may still prove indispensable for the future development of intelligent vehicle technology.

Bibliography

Primary Literature

1. Traffic Accident Causation in Europe (TRACE) (2007) Review of crash effectiveness of intelligent transport systems, Project No. 027763, Deliverable D4.1.1–D6.2, 55
2. Jones WD (2006) Safer driving in the dead of night – infrared vision systems are set to become standard in high-end cars. *IEEE Spectrum Green Technology*. <http://spectrum.ieee.org/green-tech/advanced-cars/safer-driving-in-the-dead-of-night>
3. Yendrikhovskij S (1999) Image quality: between science and fiction. In: *Proceedings of the IS&T PICS, Savannah*, pp 173–178
4. Hertel D, Betts A, Hicks R, ten Brinke M (2008) An adaptive multiple-reset CMOS wide dynamic range imager for automotive vision applications. In: *Proceedings of the 2008 IEEE intelligent vehicles symposium*, IEEE Press, Piscataway, pp 614–619
5. Hertel D (2010) Extended use of incremental signal-to-noise ratio as reliability criterion for multiple-slope wide-dynamic-range image capture. *J Electron Imaging* 19(1):011007
6. Marjukka E, Liisa H (2005) MOVE – performance based model for mesopic photometry. Helsinki University of Technology Report No. 35, ISBN 951-22-7566-X, Espoo, Finland
7. Fu C, Li C, Luo MR, Hunt RWG, Pointer MR (2007) Quantifying colour appearance for unrelated colour under photopic and mesopic vision. In: *Proceedings of the IS&T 15th color imaging conference*, Albuquerque, pp 319–324
8. Hertel D, De Locht C (2010) Utilizing near-infrared and colour information in an automotive wide-dynamic-range night vision system for driver assistance. In: Meyer G, Valldorf J (eds) *Advanced microsystems for automotive applications 2010 – smart systems for green cars and safe mobility*. Springer, Dordrecht/London/New York, pp 145–154
9. Carlson PJ, Urbanik T (2004) Validation of photometric modeling techniques for retroreflective traffic signs. *Transp Res Rec* 1862:109–118

10. Keelan BW (2002) Handbook of image quality. Dekker, New York, pp 329–399
11. Hertel D, Chang E (2007) OECF characterization of a non-linear HDR color camera for automotive applications. In: Proceedings of the IS&T 15th color imaging conference, Albuquerque, pp 244–248
12. Dierickx B (2008) Wide dynamic range: what should I care about? IEEE international solid-state circuits conference (ISSCC '08), imager design forum, wide dynamic range imaging, San Francisco
13. Williams D (2003) Debunking specsmanship: progress on ISO/TC42 standards for digital capture imaging performance. In: Proceedings of the IS&T PICS, Rochester, pp 77–81
14. Sangwine SJ, Horne REN (1998) The colour image processing handbook. Springer, New York, p 127
15. International Standard ISO 12232 (1997) Photography – electronic still picture cameras – determination of ISO speed
16. International Standard ISO 14524:1999(E) (1999) Photography – electronic still-picture cameras – methods for measuring opto-electronic conversion functions (OECFs)
17. International Standard ISO 15739:2003(E) (2003) Photography – electronic still-picture imaging – noise measurements
18. Görgens E (1987) Objective evaluation of image quality. *J Inf Rec Mater* 15:305–321
19. Kerr DA (2008) The ISO definition of the dynamic range of a digital still camera. http://dougkerr.net/Pumpkin/articles/ISO_Dynamic_range.pdf
20. Dainty JC, Shaw R (1974) Image science. Academic Press, London, pp 152–171
21. Chamberlain SG, Lee JPY (1984) A novel wide dynamic range silicon photodetector and linear imaging array. *IEEE J Solid-State Circuits* SC-19(2):41–48
22. Storm G, Henderson R, Hurwitz JED, Renshaw D, Findlater K, Purcell M (2006) Extended dynamic range from a combined linear-logarithmic CMOS image sensor. *IEEE J Solid-State Circuits* 41(9):2095–2106
23. Wang X, Wong W, Hornsey R (2006) A HDR CMOS image sensor with inpixel light-to-frequency conversion. *IEEE Trans Electron Devices* 53(12):2988–2992
24. Yang D, El Gamal A, Fowler B, Tian H (1999) A 640×512 CMOS image sensor with ultrawide dynamic range floating-point pixel-level ADC. *IEEE J Solid-State Circuits* SC-34(12):1821–1834
25. Kawahito S (2006) A sensitivity and linearity improvement on a 100 dB dynamic range image sensor using a lateral overflow integration capacitor. *IEEE J Solid-State Circuits* 41(4): 851–858
26. Decker S, McGrath RD, Brehmer K, Sodini CG (1998) A 256×256 CMOS imaging array with wide dynamic range pixels and column-parallel digital output. *IEEE J Solid-State Circuits* SC-33(12):2081–2091
27. Knoll PM (2002) A NIR based system for night vision improvement. SAE Convergence 2002, Detroit
28. Kallhammer J-E (2006) Night vision: requirements and possible roadmap for FIR and NIR systems. *SPIE* 6198, 61980 F
29. Hörter MH, Stiller C, Koelen C (2009) Hardware and software framework for automotive intelligent lighting. In: Proceedings of the 2009 IEEE intelligent vehicles symposium, Xi'an, China, p 299
30. Hamilton JF, Compton JT (2007) Processing color and panchromatic pixels. US Patent Application 20070024879
31. Honda H, Iida Y, Itoh G, Egawa Y, Seki H (2007) A novel Bayer-like WRGB color filter array for CMOS image sensors. In: Proceedings of the SPIE-IS&T electronic imaging, SPIE, vol 6492, pp 64921 J.1–64921 J.10
32. Kriesel M, Gat N (2008) True-color night vision cameras. In: Proceedings of the SPIE, vol 6540, pp 65400D.1–65400D.10
33. Frame WW (2006) One chip, low light level color camera. US 7,012,643 B2
34. Kidono K, Yoshiki N (2007) Visibility estimation under night-time conditions using a multiband camera. In: Proceedings of the 2007 IEEE intelligent vehicles conference, Istanbul, Turkey, pp 1013–1018
35. Hertel D, Marechal H, Tefera DA, Fan W, Hicks R (2009) A low-cost VIS-NIR true color night vision video system based on a wide dynamic range CMOS imager. In: Proceedings of the 2009 IEEE intelligent vehicles symposium, Xi'an, China, p 273
36. Koyama S et al (2009) A day and night MOS imager spectrally adjusted for a wide range of color temperatures. *SPIE* 7249, 72490 S
37. Catrysse PB, Wandell BA (2007) Integrated color pixel. U.S. Patent No. 7,248,297 (Filed 2001 – Issued 2007)
38. Choi B (2007) Nano-technology for color: plasmonic nano-optic filter technology for color capturing, expression, and management. In: Proceedings of the IS&T CIC15, Albuquerque
39. International Standard ISO 17321:1998 (1998) Graphic technology and photography – colour target and procedures for the colour characterization of digital still cameras (DSCs). ISO, Geneva, Switzerland
40. Debevec PE, Malik J (1997) Recovering high dynamic range radiance maps from photographs. In: Proceedings of the SIGGRAPH 97 annual conference, Los Angeles, pp 369–378
41. Mitsunga T, Nayar SK (1999) Radiometric self calibration. In: Proceeding of the IEEE computer vision and pattern recognition, Fort Collins
42. Hertel D (2010) Using spectral response measurements to optimise the colour processing of a wide dynamic range colour camera with near-infrared sensitivity. The royal photographic society, symposium on colour imaging 2010, http://www.rps-isg.org/SCI2010_programme.php
43. Scott Tyo J (2007) Displaying advanced optical imagery information for human observers. In: Proceedings of the IS&T CIC15, Albuquerque, pp 174–177
44. Ngo H, Tao L, Zhang M, Livingston A, Asari V (2005) A visibility improvement system for low vision drivers by nonlinear enhancement of fused visible and infrared video. IEEE computer society conference on computer vision and pattern recognition (CVPRW/05) – Workshops, p 25
45. Fredembach C, Süssstrunk S (2008) Colouring the near-infrared. In: Proceedings of the IS&T 16th color imaging conference, Portland, pp 176–182

46. Meylan L, Süsstrunk S (2006) High dynamic range image rendering using a retinex-based adaptive filter. *IEEE Trans Image Process* 15(9):2820–2830
47. Bertozzi M, Broggi A, Fascioli A (2000) Vision-based intelligent vehicles: state of the art and perspectives. *Robot Auton Syst* 32:1–16
48. European Commission, Information Society and Media: i2010 intelligent car initiative. http://ec.europa.eu/information_society/activities/intelligentcar/docs/icar/intelligent_car.pdf
49. Carley L (2009) News technologies for safer driving: adaptive lighting, night vision, backup cameras & blind spot detection. Import car, http://www.import-car.com/Article/60749/News_Tech.aspx
50. Kang HR, Anderson PG (1992) Neural network applications to the color scanner and printer calibrations. *J Electron Imaging* 1:125–134

Books and Reviews

- Frieser H (1975) *Photographische Informationsaufzeichnung*. Focal Press, New York
- Hunt RWG (2004) *The reproduction of colour*. Fountain Press, Tolworth, England
- Proceedings of IEEE intelligent vehicles symposium 2007, Istanbul, Turkey, 13–15 June 2007, ISBN 1-4244-1068-1
- Proceedings of IEEE intelligent vehicles symposium 2008, Eindhoven, The Netherlands, 4–6 June 2008
- Proceedings of IEEE 2009 IEEE intelligent vehicles symposium, Xi'an, China, 3–5 June 2009, ISBN 978-1-4244-3502-9
- Reinhard E, Ward G, Pattanaik S, Debevec P (2006) *High dynamic range imaging – acquisition, display, and image-based lighting*. Morgan Kaufmann Publishers, San Francisco
- Sharma G (2003) *Digital color imaging*. CRC Press, Boca Raton
- Wyszecki G, Stiles WS (2000) *Color science*. Wiley, New York

Tuberculosis, Epidemiology of

GIOVANNI SOTGIU¹, MATTEO ZIGNOL²,
MARIO C. RAVIGLIONE²

¹Epidemiology and Biostatistics Unit, Department of Biomedical Sciences, University of Sassari, Sassari, Italy

²STOP TB Department (STB), HIV/AIDS, TB & Malaria Cluster (HTM), World Health Organization, Geneva, Switzerland

Article Outline

Glossary

Definition of the Subject and Its Importance

Introduction

Basic and Descriptive Epidemiology of Tuberculosis
Etiologic Epidemiology of Tuberculosis
Tuberculosis/HIV Coinfection
Drug-Resistant Tuberculosis
Control of Tuberculosis
Global Epidemiological Situation
Future Directions
Bibliography

Glossary

Acquired immunodeficiency syndrome (AIDS) Clinical syndrome caused by the Human Immunodeficiency Virus (HIV). Its pathogenesis is related to a qualitative and quantitative impairment of the immune system, particularly a reduction of the CD4+cell count (surrogate marker of the disease). After an average of 10 years if untreated, HIV + individuals can develop opportunistic diseases (i.e., infections and neoplasias rarely detected in immunocompetent subjects). The natural history of the disease can be dramatically modified with administration of combination therapy composed of at least three antiretroviral (ARV) drugs.

Human immunodeficiency virus (HIV) Virus that causes Acquired Immunodeficiency Syndrome (AIDS). It belongs to *Retroviridae* family and was discovered in 1983 by Luc Montagnier and Robert Gallo. It is transmitted mainly through sexual intercourse, exchange of contaminated syringes among intravenous drug users, and contaminated blood transfusion. HIV-1 is the type most frequently detected worldwide.

Incidence Rate describing the number of new cases of disease occurring within a unit of time in a defined cohort at risk of disease (expressed in cases per 100,000 population per year).

Latent tuberculosis infection (LTBI) Infection caused by *Mycobacterium tuberculosis* transmitted mainly through the air. Clinical and/or radiological signs of latent tuberculosis infection cannot be detected in the majority of the cases. The infection can be presumptively diagnosed by a positive tuberculin skin testing and/or a positive interferon- γ release assay (IGRA), being able to identify a persistent adaptive immunological reactivity against mycobacterial antigens. In most individuals

mycobacteria can be eliminated through chemoprophylaxis. It is estimated that one third of the human population is infected by *Mycobacterium tuberculosis* worldwide.

Mortality Rate describing the number of deaths from a disease occurring within a unit of time in a defined cohort at risk of death (expressed in deaths per 100,000 population per year).

Mycobacterium tuberculosis Bacterium that causes tuberculosis, discovered by Robert Koch in 1882. It is genetically closely related to other mycobacteria with which it forms a complex (*Mycobacterium africanum*, frequently detected in Western Africa, and *Mycobacterium bovis*, frequently detected in the past in cows and transmitted to human beings through unpasteurized milk).

Prevalence Number of cases of disease in a defined population at a specific point in time; it is mainly presented as a relative frequency (i.e., proportion usually expressed per 100,000 population).

Tuberculosis Infectious disease caused by *Mycobacterium tuberculosis*. It usually involves the lungs (pulmonary tuberculosis) but can also affect other organs (i.e., kidneys, central nervous system, lymph nodes, bones, etc.: extrapulmonary tuberculosis). Pulmonary tuberculosis, which is the most frequent clinical form, can be classified as smear-positive or smear-negative according to the result of the sputum bacteriological examination. The former is a major public health problem being highly contagious. Only a few individuals develop tuberculosis after a mycobacterial infection, and most of them soon after infection: it is estimated that the lifetime risk is 5–10% in HIV-negatives and 5–15% yearly in HIV-positives.

Definition of the Subject and Its Importance

Tuberculosis is a disease caused by bacilli belonging to the *Mycobacterium tuberculosis* complex, which includes the species *Mycobacterium tuberculosis* (most frequently detected in human beings), *Mycobacterium africanum*, and *Mycobacterium bovis*. It is primarily an airborne disease. Mycobacteria enter into the human airways through the inhalation of droplet nuclei, i.e., particles containing mycobacteria aerosolized by coughing, sneezing, talking, or singing. Other rare, recognized ways of transmission are the ingestion of

cow milk contaminated by *Mycobacterium bovis* (frequent route of transmission in the past), cutaneous inoculation in laboratory workers and pathologists, and sexual intercourse [1].

Tuberculosis is a major public health issue worldwide. Together with HIV/AIDS, it is one of the main infectious killers of individuals in their productive years in low-income countries [2]. However, tuberculosis is a highly curable disease if diagnosed and treated with combination chemotherapy. As a result of the global implementation of the World Health Organization's (WHO) STOP TB Strategy, during the period 1995–2009, 41 million tuberculosis patients were successfully treated and 6 million deaths averted compared to what would have happened if current standards were not implemented [3, 4].

Introduction

The epidemiology of tuberculosis studies the dynamics and interactions between *Mycobacterium tuberculosis* and human beings in a specific environment. Furthermore, it qualitatively and quantitatively evaluates all the covariates (for instance, the administration of antituberculosis drugs) that could interfere with the outcomes of the natural history of this interaction (i.e., exposure, infection, disease, and death) [5].

Basic and Descriptive Epidemiology of Tuberculosis

The global burden of tuberculosis is measured using epidemiological indicators (incidence, prevalence, mortality, and case fatality). Estimates of these indicators are computed yearly by the World Health Organization which collates information obtained from surveillance systems (notification system and mortality registries), special epidemiological studies (surveys of tuberculosis prevalence and in-depth analyses of surveillance data), and experts' opinion [6].

Epidemiological Indicators

Tuberculosis incidence: number of new cases of tuberculosis occurring within a specific time period (usually 1 year) in a defined cohort. It is usually presented as a rate per 100,000 inhabitants and describes the probability of developing tuberculosis in a specific time period [6, 7].

Tuberculosis prevalence: number of cases of tuberculosis in a defined population at a specific point in time. It is presented as an absolute or a relative (usually per 100,000 inhabitants) frequency and can be considered the product of the incidence of tuberculosis by the duration of the disease. This indicator assesses the global, national, regional, and local burden of tuberculosis.

Tuberculosis mortality: number of deaths from tuberculosis occurring within a specific time period (usually 1 year) in a cohort of individuals with tuberculosis disease. It is often presented as a rate per 100,000 people and describes the probability of dying from tuberculosis in a specific time period.

Tuberculosis case fatality: number of deaths from tuberculosis occurring within a specific time period (usually 1 year) in a defined cohort of individuals with tuberculosis. It is presented as a percentage.

Incidence of sputum smear-positive pulmonary tuberculosis cases: number of new sputum smear-positive cases of pulmonary tuberculosis occurring within a specific time period (usually 1 year) in a defined cohort. It is usually presented as a rate per 100,000 population. It identifies the most contagious tuberculosis cohort, that is, the most important source of infection.

Incidence of sputum culture-positive pulmonary tuberculosis cases: number of new sputum culture-positive cases of pulmonary tuberculosis occurring within a specific time period (usually 1 year) in a defined cohort. It is usually presented as a rate per 100,000 population.

Etiologic Epidemiology of Tuberculosis

Natural History

The dynamics of the probabilistic model of tuberculosis are complex and the associated covariates, as well as their quantitative effects, are not always known. The potential interaction between a human being and the strains of *Mycobacterium tuberculosis* could result in the following non-deterministic outcomes: subclinical infection, pulmonary and/or extrapulmonary disease, and death. The incidence of the above mentioned outcomes is regulated by the occurrence and the additive and/or the synergistic combination of multiple risk factors. Several environmental, bacterial, and human

risk factors have been identified. Some of them could play an important role since the first crucial event, that is, the exposure to a source of mycobacteria [1, 5, 8].

Exposure

A critical exposure to transmission of mycobacteria from a contagious patient to a susceptible individual could be a close physical contact or permanence in a small room with limited ventilation [5].

The most important variable significantly increasing the likelihood of exposure to a source of *Mycobacterium tuberculosis* is the prevalence of contagious patients in a specific setting. This is estimated as the product of the incidence of infectious individuals by the duration of their infectiousness. The number of incident cases greatly varies between settings and depends on numerous factors.

Duration of infectiousness is strictly related to the capacity of the health system for early detection and adequate treatment of an index case [5, 8]. After a microbiological diagnosis, which should include drug-sensitivity testing, proper anti-tuberculosis treatment should be started without delay [3]. Correct drug combinations and dosages should be used to avoid sub-optimal anti-bacterial activity [5, 9, 10]. Strict clinical follow-up of patients in nosocomial or community settings should be undertaken to ensure adherence to anti-tuberculosis medications for the full duration of treatment. Compliance could be compromised by the long duration of therapy and/or the emergence of adverse events. The capacity of a patient to adhere to antituberculosis therapy depends on factors related to the health system, socioeconomic conditions of the individual, type of therapy, and the disease and patient's characteristics. The concept of early case detection and immediate treatment is supported by the evidence that on average 30–40% of the close contacts of a sputum smear-positive pulmonary case are estimated to be infected at the time of diagnosis. The public health consequences of a diagnostic and/or therapeutic delay are dramatic [5].

It is clear that quality and quantity of human relations modify the risk of critical exposures and, consequently, the risk of acquiring the infection [5]. In particular, population density is a critical element: the higher the population density, the greater the risk of

interactions. High population density can be experienced in urban areas or in household crowding associated with poor housing [5, 8]. Climatic condition is an additional environmental variable that plays a relevant role in the likelihood of exposure. In cold climates, people tend to spend more time in indoor activities, increasing the risk of exposure because of the higher chances of close contact. On the other side, in warm climates people tend to spend more time outdoors, diminishing frequency and duration of case–contact interactions per unit of space. At the same time natural indoor ventilation and the effect of the solar ultraviolet rays that kill mycobacteria contribute to control the spreading of infection [5].

Other relevant non-environmental modifiers of the risk of exposure are sex and age. In several low-income countries and social groups, males and females have different opportunities of social contact and therefore a different risk of exposure. For instance, in some countries females are excluded from public activities and spend most of their time indoors. The median age of tuberculosis cases, and consequently of contagious sources of infection, is considerably different between low- and high-income countries, with elderly people representing the majority of the autochthonous patients in industrialized areas and young adults counting for the largest proportion of patients in low-income countries [5, 11–13].

Infection

The risk of infection by *Mycobacterium tuberculosis* is directly associated to the probability of exposure to infectious particles produced by a tuberculosis patient through coughing, sneezing, talking, or singing [1, 5, 8, 14]. Such risk is related to the concentration of contagious droplet nuclei containing mycobacteria and to the time of exposure [1, 5].

Infectious particles must be suspended in the air in order to be inhaled by a contact of a tuberculosis patient. The speed of falling to the ground is directly correlated to the square of their diameter. However, the tendency of liquid particles to evaporate and, consequently, to reduce their diameter, increases the speed of their descent. The effect of evaporation on very large liquid particles is less important. Furthermore, humidity and temperature could influence the

evaporation: high humidity hinders the evaporation of infectious particles and, indirectly, increases the speed of dropping to the ground. Overall, the effect of the diameter of the particle is more influential than that of humidity [5].

Droplet nuclei should be 1–5 μm large to be inhaled and retained in the pulmonary alveoli. It has been demonstrated that a diameter $>5 \mu\text{m}$ increases the probability of being entrapped in the upper airways through the continuous movement of both the vibrissae and the muco-ciliary system. On the other hand, contagious droplets whose diameter is $<1 \mu\text{m}$ can arrive to the alveolar spaces, but the probability of retention in the peripheral pulmonary alveoli is very low [1, 5].

A crucial variable that can dramatically change the air density of contagious droplet nuclei in a specific indoor environment (for instance, in a nosocomial setting) is the natural, mixed-mode, or mechanical ventilation [15]. Several authors described transmission of mycobacterial strains in health-care settings with flawed or missing ventilation systems both in low- and high-income countries. The typology of the ventilation system should be chosen taking into account several variables: the building structure, outdoor climatic conditions and air quality, purchase and maintenance costs, and national and local regulations. However, modifiers should not alter the main principle of a ventilation system that must promote an airflow direction from the infectious source to the air exhaust area [5, 15]. Ventilation rate can be evaluated measuring the volume of the space (i.e., air changes per hour – ACH) or the number of individuals in a space (i.e., liters/second/person): every individual should have a definite supply of fresh air to dilute the concentration of mycobacteria. The World Health Organization recommends more than 11 air changes per hour for an isolation room of 24 m^3 (i.e., at least 80 l/s/person). Natural ventilation may be improved by increasing the size of windows and positioning them on opposite walls. Well-designed fans can improve the airflow direction when natural ventilation rates are not adequate (mixed-mode ventilation). However, continuous maintenance of those fans is essential. In some circumstances it may be necessary to use upper room or shielded ultraviolet germicidal irradiation (UVGI) devices. In settings with frequent climatic changes or

high risk of transmission of multi-drug resistant mycobacterial strains, UVGI devices could be helpful in addition to a mechanical ventilation system.

Several measures should be implemented in household settings to reduce the risk of transmission: improvement of natural ventilation, education on cough etiquette, and respiratory hygiene. Furthermore, a tuberculosis patient should sleep isolated in a well-ventilated room or, if his/her clinical conditions are favorable, should spend as much time as possible outdoors.

Personal protective equipment may considerably reduce the possibility of infection, especially when the ventilation system is weak. In this regard only high-efficiency particulate air-filter respirators filtering out droplet nuclei sized 1–5 μm can be protective. Particulate respirators meeting or exceeding the N95 standards set by the United States Centers for Disease Control and Prevention/National Institute for Occupational Safety and Health (CDC/NIOSH) or the CE certified FFP2 standards are recommended for health-care workers (particularly during the management of multi-drug resistant tuberculosis cases and aerosol-generating procedures) and visitors. Health education on the importance of respirators in communities at risk should be carried out in order to avoid stigma, and continuous training of health-care providers on the correct use (fit testing) and on indications of respirators should be organized. Surgical masks can be worn by contacts of a patient but do not protect from exposure [15]. The probability of being infected can be decreased if the source of droplet nuclei is isolated or if both mouth and nose are covered.

The most efficient measure to stop transmission is to rapidly detect and adequately treat a tuberculosis case [3, 5, 8, 15]. Patients are normally not infectious after 2–3 weeks of effective treatment. However, not all tuberculosis patients are equally infectious and there is a correlation between the number of mycobacteria in a milliliter (ml) of sputum and the efficiency of transmission. At least 5,000 mycobacteria in 1 ml of sputum are necessary to classify a sputum smear examination as positive.

Several studies demonstrated that sputum smear-positive patients are significantly more contagious than those who are sputum smear-negative and culture-positive. Furthermore, the latter are not much more

infectious than tuberculosis patients who have a negative smear and culture [5, 8].

Although sputum smear-positive patients are the most contagious population, those who are sputum smear-negative need equal attention, as 17% of mycobacterial transmissions, in some industrialized settings, are due to this group of patients [5].

According to estimates computed in the pre-antibiotic era and used to model the tuberculosis epidemic, an undiagnosed sputum smear-positive patient can infect 10–12 contacts annually. Since after 2 years, the probability of death or spontaneous conversion to sputum smear-negativity is high, the average number of contacts infected by a single smear-positive case was estimated to be 20–24 [5, 8, 16]. The delay of diagnosis and/or of treatment of just 2–3 months could result in infection of several contacts [5].

Numerous factors determine patient's delay to diagnosis: these include alcohol or substance abuse, low income, difficult access to health care facilities, disabilities, and beliefs about tuberculosis. On the other hand, other concomitant pulmonary diseases, sputum smear-negative or extra-pulmonary tuberculosis, inexperience of health care workers, weak health care infrastructure, and absence of respiratory symptoms are all factors associated with healthcare delay [17, 18].

The annual burden of the most relevant source of tuberculosis infections is assessed through an epidemiological indicator: the incidence of sputum smear-positive pulmonary tuberculosis cases. This describes the number of new sputum smear-positive cases of pulmonary tuberculosis occurring within 1 year in a defined population at risk of tuberculosis [5, 6, 8]. The average annual risk of latent tuberculosis infection is an important epidemiological indicator that assesses the probability of becoming infected with *Mycobacterium tuberculosis* in a period of 1 year [5, 8].

Clinical and/or radiological signs of latent tuberculosis infection cannot be detected in the majority of the cases. However, it can only be presumptively diagnosed through a tuberculin skin testing (TST) and/or an interferon- γ release assay (IGRA), which identifies a persistent adaptive immunological reactivity against mycobacterial antigens [14, 19–22]. The tuberculin skin test, performed with an intradermal injection of purified protein derivative (PPD) obtained from *Mycobacterium*

tuberculosis culture filtrate, has been considered as the gold standard diagnostic tool for identifying individuals latently infected with *Mycobacterium tuberculosis*. The main limitation of this technique is the low specificity, due to antigenic similarities between PPD, Bacille Calmette-Guerin (BCG) strains, and nontuberculous mycobacterial strains. Furthermore, its application and interpretation significantly depend on the health worker responsible for the test [14, 19, 22].

Interferon Gamma Release Assays (IGRA), available in two different commercial forms, that is, QuantiFERON-TB Gold In Tube (QFT-IT; Cellestis Ltd., Chadstone, Australia) and T-SPOT.TB (Oxford Immunotec, Abingdon, UK), are tools for the in vitro diagnosis of latent tuberculosis infection. They detect cellular immune reactivity toward *Mycobacterium tuberculosis*-specific antigens (ESAT-6, CFP-10, and TB7.7). As such, contrary to PPD, they can discriminate between people previously immunized with Bacille Calmette-Guerin vaccination or those infected by non-tuberculous mycobacteria from tuberculosis patients [14, 19–22].

The average annual risk of tuberculosis infection can be estimated with direct or indirect methods:

- The direct evaluation assesses the proportion of individuals who convert to tuberculin skin test in a defined period of time. The main limitations are related to the sample size, which should be large enough in order to obtain an adequate statistical power, and to the boosting phenomenon (an apparent conversion related to repeated tests).
- The indirect evaluation relies on the relationship between the prevalence of tuberculosis infection in a specific time-point and the average annual risks of infection.

The probability of being infected is estimated to be very low in high-income countries (from 0.1% to 0.01%) especially among the young. On the other hand, the prevalence of infection is high in most developing countries, due to ongoing transmission [5, 8]. Overall, one third of the global population is estimated to be infected by *Mycobacterium tuberculosis* [5, 6, 8].

Disease

Active tuberculosis is the outcome of infection with *Mycobacterium tuberculosis* [1, 5, 8, 23]. As for exposure

and infection, the likelihood of clinical tuberculosis is strictly dependent on risk factors, the impact of which is not only related to their effect on the pathogenesis of the disease but also to their prevalence in the general population [5, 8].

After the establishment of tuberculosis infection the probability of developing a disease is higher in the first 12–24 months after the entry of *Mycobacterium tuberculosis* into the human body, and decreases as time elapses from the infection. It is generally estimated that the risk of disease is about 5% in the first 1–2 years after infection, and up to 5% in the subsequent years [1, 5, 8]. The age of primary infection is an important determinant of the cumulative risk, as this is the highest in young individuals (from 35% to 50% in children aged less than 15 years and those who are in close contact with a sputum smear-positive patient) [5, 8].

The key pathogenic factor favoring the development of tuberculosis is the quantitative and/or functional deficiency of the innate and/or adaptive immune system, which is responsible for the control of tuberculosis infection [2, 5, 8, 23]. The host cellular and molecular mechanisms triggered by mycobacterial infections are only partially known. After reaching the lower respiratory tract, the mycobacteria grow, partially controlled for the first few weeks by nonspecific immunity. CD4⁺ lymphocytes producing interferon- γ are recruited after 2 weeks but their number is not sufficient to activate the pool of alveolar macrophages. Their priming in the draining lymph nodes occurs only after the first 7–10 days because mycobacteria are initially phagocytosed by non-motile phagocytic cells. T-cell priming and accumulation could additionally be delayed through the activation of regulatory T lymphocytes. Continuous activation of T lymphocytes is vital to control mycobacteria in the granuloma; nevertheless, several experimental data demonstrated that once exposed to chronic mycobacterial stimulation, T lymphocytes can experience a functional collapse [1, 23]. Numerous medical conditions can modify one or more components of the immune response against *Mycobacterium tuberculosis*, compromising the control of the infection, and, then, increasing the probability of development of disease [1, 5, 8]. For these reasons and due to its global burden, HIV/AIDS is recognized as the most important risk

factor for the development of tuberculosis [1, 2, 5, 8, 24, 25]. Numerous studies demonstrated that the lower the CD4+ lymphocytes counts the higher the risk of developing pulmonary as well as extrapulmonary disease. The qualitative and quantitative role of CD4+ lymphocytes in the adaptive immune response against mycobacteria is crucial, particularly their interaction with macrophages, which need to be activated in order to be effective. The risk of development of tuberculosis disease is 5–15% per year if the mycobacterial infection precedes HIV transmission. On the contrary, such risk could be significantly higher if mycobacterial infection develops in a seriously immunocompromised host. Generally, the probability of developing tuberculosis is 20–37-fold higher in HIV-infected compared to HIV-negative individuals. The relative risk is 20.6 in countries with a generalized HIV epidemic, 26.7 in countries with concentrated epidemics, and 36.7 in low HIV-prevalence countries [1, 2, 5, 8, 26].

Another medical condition that could compromise the immune key players against mycobacterial strains is diabetes mellitus. The estimated prevalence of diabetes is 180 million individuals globally, a figure expected to increase to 360 million by 2030 [5, 27]. The probability of developing tuberculosis is 1.5–8 times higher among individuals with diabetes compared to healthy subjects. No significant differences in terms of risk of disease have been detected between low- and high-income countries [28–32]. Experimental studies showed that hyperglycemic mice have higher mycobacterial loads, impaired production of interferon- γ and interleukin-12 together with an impaired T helper 1 response. Leukocyte bactericidal activity is reduced in diabetic patients compared to healthy individuals. Diabetic patients have an altered chemotaxis and oxidative killing of neutrophils. Furthermore, impaired immunity is frequently associated with other relevant risk factors like chronic renal failure, malnutrition, and pulmonary microangiopathy [28, 33, 34]. Nevertheless, several epidemiological analytical studies assessing the relative risk or the odds ratio of diabetic patients are biased because routine data sources were used, which do not allow controlling for all confounders. The population-attributable risk for diabetes is similar to that of HIV/AIDS; although the relative risk related to HIV/AIDS is highest, ranging from 6.5 to 26 (i.e., 2–9 times higher

than the relative risk of diabetes), it is less prevalent than diabetes [28–33, 35].

Chronic renal failure and the hemodialytic treatment increase the risk of developing tuberculosis from 6.9 to 52.5 times compared to the general population [5, 36]. Some epidemiological studies highlight a greater incidence of tuberculosis in the first year of dialysis due to the occurrence of a severe immune depression. It is difficult to understand the role of chronic renal failure because the majority of the studies enrolled principally hemodialytic patients. The crucial pathogenetic feature is the impaired immune response to mycobacterial isolates, in particular type 1 helper T-cell responses involving interleukin-12 and interferon- γ production and co-stimulatory function of antigen-presenting cells. In the final stages of a chronic renal failure, a high rate of anergy to mycobacterial antigens administered intracutaneously has been documented (i.e., from 32% to 40%). Other relevant elements are the continual inflammatory phase of monocytes/macrophages due to the uremia and to the dialysis, malnutrition (mainly related to vitamin D intake), and hyperparathyroidism [5, 36–38].

Silicosis increases the likelihood of pulmonary tuberculosis from 2.8 to 39 times if compared to healthy controls whereas the probability of developing extra-pulmonary tuberculosis is 3.7-fold greater (the pleural form is the most common form, 61% of cases) [1, 5, 24, 39–41]. The average time elapsing between the diagnosis of silicosis and the development of tuberculosis is 6.8 years. The incidence of tuberculosis seems to be proportional to the severity of silicosis and the intensity of exposure to crystalline silica dust. Occupational activities in the mines seem to modify considerably the risk of developing tuberculosis: in particular, drilling has been associated with a greater intensity of exposure. Some studies identified a higher relative risk (i.e., 1.1–4.0) of developing tuberculosis in miners exposed to silica but without silicosis. These epidemiological findings support the results of experimental studies demonstrating that silica modifies the pulmonary immune response, and impairs the metabolism and the function of pulmonary macrophages till their apoptosis after long exposure [5, 24, 39–41]. The incidence of silica-related tuberculosis has been increased by the increased prevalence of HIV infection in resource-limited countries. The incidence of

pulmonary tuberculosis in South African gold miners is 3,000 per 100,000 population. For this reason, several authors strongly recommend the treatment for latent tuberculosis infection in individuals with and without silicosis (especially for HIV-infected people) [5, 39–41].

Smoking is another relevant modifier, being associated to a relative risk of developing tuberculosis ranging from 2.3 to 2.7 [1, 5, 24, 42, 43]. Furthermore, it increases the risk of infection (relative risk of 1.7). The relative risk for latent tuberculosis infection and disease is not independent: the increase of the risk of infection directly increases the proportion of individuals at risk of disease. Therefore, the independent relative risk for tuberculosis in an infected population is obtained computing the ratio of the two relative risks (i.e., 1.4–1.6). Smoking reduces adaptive immune responses, decreasing CD4+ cell counts, altering macrophages, and mechanically disrupting the cilia in the upper respiratory airways. A dose–response effect has been documented: the likelihood of tuberculosis is directly linked to the number of cigarettes smoked daily.

Being treated with immunosuppressive drugs could increase the risk of tuberculosis, as recently demonstrated with the anti-TNF α therapies for the treatment of chronic inflammatory diseases (for instance, rheumatoid arthritis, psoriasis and psoriatic arthritis, ankylosing spondylitis, juvenile idiopathic arthritis, and inflammatory bowel disease) [44]. TNF is a protein crucial for granuloma initiation and stability in the lungs. It fosters the phagocytic and killing activities of the alveolar macrophages and favors the gathering of cells at the inflammatory site inducing adhesion molecules. Four monoclonal anti-TNF antibodies are currently available and are used for the treatment of chronic inflammatory diseases: adalimumab, certolizumab pegol, golimumab, and infliximab. The risk of developing tuberculosis is amplified by 1.6–25.1 times and varies depending on the clinical setting and type of anti-TNF therapy prescribed [44–46].

The role of a corticosteroid treatment in the pathogenesis of tuberculosis is debated [5]. Daily dose of 10 mg or higher of corticosteroids, prescribed for a few days, appears to not be associated with an increased risk of developing tuberculosis but experimental studies on rabbits documented their role in increasing the risk of disease.

Numerous studies evaluated the impact of several tumors as risk factors for tuberculosis, but only lymphomas, pulmonary neoplasia, and head/neck cancers seem associated to the development of the disease [5]. Other relevant diseases and medical events that could increase the risk of tuberculosis are malnutrition, substance abuse, gastrectomy, and jejunoileal by-pass [5, 24]. Intravenous drug and alcohol abuse could increase the risk of tuberculosis by compromising the immune system. Malnutrition, meat and fish deficiencies, or a vegetarian diet are epidemiologically associated to increased probability of tuberculosis [5, 24].

Mortality

The wide availability of anti-tuberculosis drugs has decreased mortality from tuberculosis and case fatality in developed countries. In the pre-antibiotic era, 30% and 66% of patients with tuberculosis died after 1 and 5 years from the notification of the disease, respectively [5, 8]. The most important prognostic variables are the site and type of tuberculosis, the early and adequate administration of anti-tuberculosis treatment, and adherence to treatment [3, 5, 8]. The majority of deaths caused by tuberculosis are attributable to sputum smear-positive tuberculosis [5, 8].

Patient's delay and/or health-care system's delay are associated to increased mortality. Patient's delay can be reduced with health education measures, decreasing tuberculosis stigma, and improving the health-care network. The delay of the health-care system could be related to several factors, particularly poor availability of diagnostic means and misdiagnosis; furthermore, negative prognostic health-care-associated factors are poor accessibility to antituberculosis drugs, incorrect prescription of anti-tuberculosis regimens in terms of dosages, time of exposure, and drug susceptibility of the infecting mycobacterial strains.

Tuberculosis/HIV Coinfection

HIV infection has dramatically changed the epidemiology of tuberculosis in high- and low-income countries, mostly in sub-Saharan Africa, Southeast Asia, and Eastern Europe. It was estimated that tuberculosis incidence in 2008, if compared with the rate computed in 1990, increased by 11% due to the HIV pandemic. In sub-Saharan Africa, tuberculosis mortality has tripled

in the last two decades and 29% of tuberculosis-related deaths are associated to HIV infection [1, 2, 5, 8, 47, 48]. The increase in tuberculosis notification rates mirrored the increase in the HIV prevalence with a delay of 4–7 years [2, 48–50].

By the end of 2008, 33.2 million individuals were estimated to be living with HIV. Incidence and mortality in 2007 were 2.7 million and 2.1 million, respectively [48, 49]. Tuberculosis is estimated to be the cause of 26% of the deaths in people living with HIV. The majority of TB/HIV patients is located in sub-Saharan Africa but in numerous high HIV prevalence areas only 20% HIV-infected individuals know their HIV status. The infection is principally transmitted through sexual relationships, exchange of contaminated syringes among intra-venous drug users, and transfusion of infected blood. The risk of developing tuberculosis ranges from 3% to 15% per year in HIV coinfecting patients, while the cumulative lifelong risk in non-HIV-infected individuals is 10%. However, surveys carried out in high HIV prevalence areas estimated an annual probability of developing tuberculosis up to 30% in advanced immunocompromised patients [1, 2, 5, 8, 48–50].

The risk of developing tuberculosis is directly correlated to the degree of immunodeficiency, particularly with the quantitative impairment of CD4+ cell counts, which are considered a relevant surrogate marker of infection together with the HIV-viral load. For this reason tuberculosis is the main opportunistic disease and cause of death in HIV + individuals globally [1, 2, 5, 8, 48, 50].

It is crucial to detect HIV-positive individuals and to start antiretroviral therapy (ART) to decrease the probability of developing tuberculosis and, consequently, the probability of transmitting mycobacterial strains to susceptible individuals. The risk of developing tuberculosis is reduced by 70–90% in patients treated with ART if compared with untreated HIV-positive individuals. In 2008, more than four million individuals were treated with antiretroviral drugs [1, 2, 5, 8, 48–54].

The role of HIV infection on tuberculosis incidence is strictly related to the prevalence of latent tuberculosis infection in the population, particularly in those aged 15–54 years. In industrialized countries, where the prevalence of the mycobacterial infection is low in all

age groups, incidence and prevalence of tuberculosis in HIV + individuals are also low. However, in some metropolitan areas an epidemiological scenario comparable to that described in some low-income countries, with high proportions of coinfecting individuals has been documented [2, 5, 8, 47, 48].

It is estimated that 30% of HIV + individuals are infected by mycobacterial strains worldwide, with a wide variability from 14% in Europe to 46% in Southeast Asia.

A series of tuberculosis/HIV collaborative activities have been developed to address tuberculosis/HIV coinfection (Table 1) [2, 3, 48, 49, 51, 55, 56]:

- Establishment of a tuberculosis/HIV coordinating body to plan and carry out common activities, such as monitoring and evaluation of the epidemiological indicators and clinical outcomes.
- Surveillance of the prevalence of HIV infection in patients with tuberculosis.
- Intensified tuberculosis case finding aimed at identifying tuberculosis symptoms and clinical signs in HIV-positive individuals attending specialized

Tuberculosis, Epidemiology of. Table 1 Tuberculosis/HIV collaborative activities [3]

1.	<i>Establishment of the mechanisms for collaboration</i> • Set up a coordinating body for tuberculosis/HIV collaborative activities that could be effective at all levels • Carry out surveillance of HIV-prevalence among tuberculosis cases • Carry out joint tuberculosis/HIV planning • Conduct monitoring and evaluation
2.	<i>To decrease the burden of tuberculosis in people with HIV/AIDS</i> • Implement intensified tuberculosis case finding • Introduce isoniazid preventive therapy • Ensure the control of latent tuberculosis infection in health-care and congregate settings
3.	<i>To decrease the burden of HIV in patients with tuberculosis</i> • Provide HIV testing and counseling • Introduce HIV-prevention methods • Introduce cotrimoxazole preventive therapy • Ensure care and support for people with HIV/AIDS • Introduce antiretroviral therapy

health-care settings (HIV clinics, sexually transmitted diseases, clinics, etc.).

- Isoniazid preventive therapy (IPT) in HIV-positive individuals latently infected by mycobacterial strains. In these cases, it is strongly recommended to microbiologically exclude tuberculosis before the administration of the anti-tuberculosis drug.
- Infection control measures in health care and congregate settings to avoid the spread of mycobacterial strains.
- HIV counseling together with HIV rapid testing in tuberculosis patients.
- Introduction of HIV-prevention methods.
- Cotrimoxazole preventive therapy.
- Wide availability of antiretroviral (ARV) drugs.

Delivery of ART to HIV-positive patients with tuberculosis is deemed one of the most important measures for a successful DOTS strategy as it can prevent the shift from latent tuberculosis infection to active tuberculosis following the immunological recovery. However, several issues can hamper the efficacy of ART in tuberculosis patients: pharmacokinetic interactions between anti-HIV drugs and anti-tuberculosis drugs and immune-reconstitution syndrome (IRIS) could increase morbidity and mortality, for instance, decreasing the adherence and, then, the clinical efficacy of the specific treatment regimens [2, 48, 50–54].

The combination of three of the above mentioned tuberculosis/HIV activities, that is, Intensified Case Finding, Isoniazid Preventive Therapy and Infection Control measures, known as the “Three I’s”, can have a significant role in the prevention of mycobacterial infections and tuberculosis, and in the early identification of tuberculosis cases [2, 3, 55, 56].

Intensified Case Finding helps to rapidly identify tuberculosis suspects using questionnaires focused on symptoms and clinical signs of tuberculosis. Individuals with or at high risk of HIV infection or in congregate settings (such as mines, prisons, etc.) are regularly screened and counseled by HIV services and/or by community-based organizations supporting HIV-positive patients.

Isoniazid Preventive Therapy can be administered to HIV-positives without tuberculosis, reducing the risk of developing the disease by 33–67% for up to 4 years. The criteria for treatment include HIV-positive

individuals living in areas with a proportion of latent tuberculosis infection above 30% or HIV-positives with documented latent tuberculosis infection or exposure to a contagious tuberculosis index case. The efficacy of its combination with the ART in preventing the development of tuberculosis has been proven.

Tuberculosis Infection Control measures can prevent the spread of mycobacterial strains to vulnerable patients, health-care workers, the community, and those living in congregate settings.

However, slow implementation of the “Three I’s” has been documented because of the difficulty faced by some national programs to integrate these components in the framework of HIV care. Some national tuberculosis programs do not support the implementation of some activities (for instance, the use of isoniazid preventive therapy) and gaps in policy and operational guidance remain a major obstacle in some settings. Infection control still represents a serious challenge at national-, regional-, and facility-level for the absence of a coordinating body or for multiple coordinating bodies, lack of technical expertise, trained health care workers, and laboratory biosafety. Furthermore, inadequate diagnostic tools for latent tuberculosis infection and tuberculosis have hampered the implementation of intensified tuberculosis case finding and the delivery of isoniazid preventive therapy predominantly in resource-limited settings [2, 48, 55, 56]. Nevertheless, it was estimated that 1.4 million tuberculosis cases were tested globally for HIV infection in 2008, which is a 64-fold increase if compared with the estimates in 2002. The increase was significant in sub-Saharan Africa (from 4% to 45% in 2004 and 2008, respectively). A significant improvement has been documented in the number of HIV-infected people screened for tuberculosis: from 600,000 in 2007 to 1.4 million in 2008. Improvements were also recently recorded on the provision of preventive and therapeutic drugs, particularly trimethoprim-sulfamethoxazole preventive treatment (200,000 HIV-positives with tuberculosis in 2008), isoniazid preventive therapy (50,000 HIV-positives in 2008), and ART (100,000 HIV-positives with tuberculosis in 2008) [2, 48, 49].

All countries with a high HIV prevalence should be committed to rapidly implement and scale-up all tuberculosis/HIV collaborative activities.

Future investments in the translational research aimed at identifying new diagnostics and therapeutics as well as international and national political commitment could improve the epidemiology of tuberculosis/HIV coinfection [2, 5, 8, 48, 50, 53].

Drug-Resistant Tuberculosis

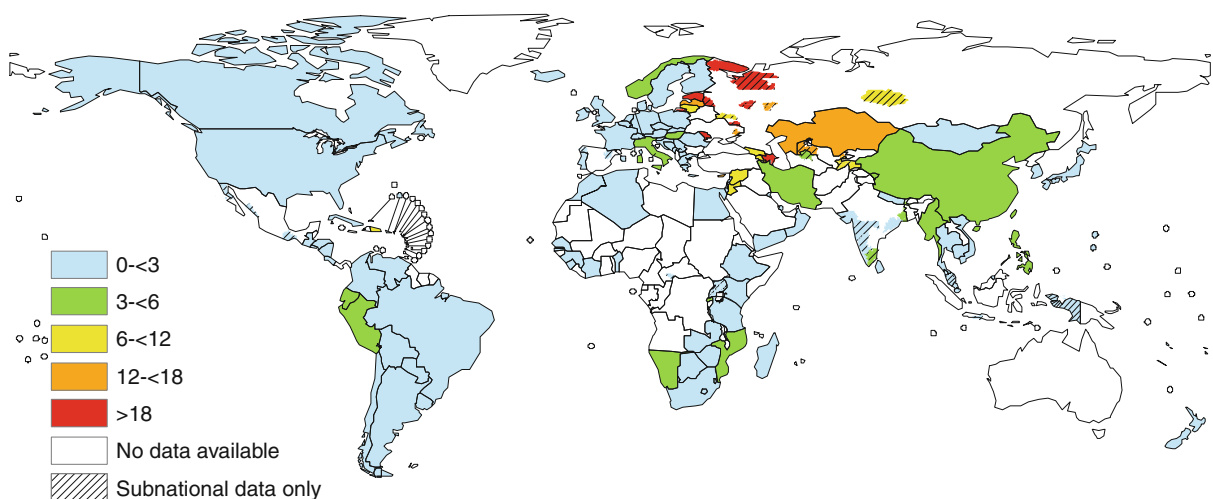
Development of resistance to anti-tuberculosis drugs has been documented since the early years of introduction of chemotherapy for the treatment of tuberculosis. The large majority of patients treated with streptomycin in the first Medical Research Council randomized clinical trial in the 1940s acquired resistance to that drug [57]. The spread of drug-resistant strains was soon recognized, and a survey of clinics in England in the 1950s found that over 5% of patients with tuberculosis without a history of previous treatment had strains resistant to at least one of the three major drugs in use at that time [58]. If even once the three effective drugs are used in combination, the development of drug resistance is theoretically impossible [59]. However, despite the introduction of combination regimens throughout the world many years ago, drug resistance has been progressively documented [60].

In the early 1990s, several reports were published on the emergence of multi-drug resistant TB (MDR-TB), but the populations surveyed were not comparable,

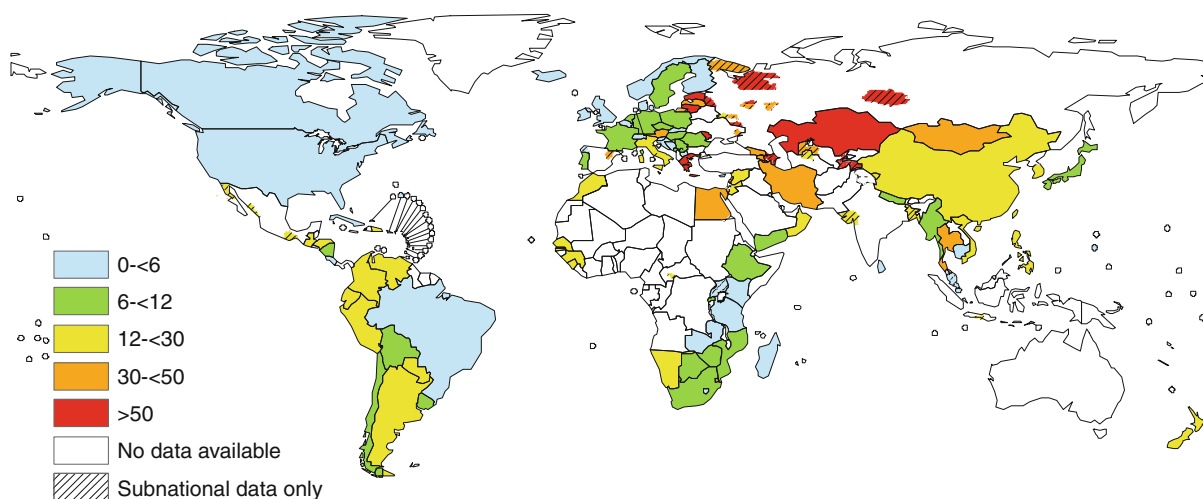
and methods used to quantify the problem were not standardized thus making it difficult to estimate the global magnitude of drug-resistant TB [61–64].

In 1994, the Global Project on Anti-tuberculosis Drug Resistance Surveillance was established in order to estimate the global burden of drug-resistant TB worldwide using standardized methodologies so that data could be compared across and within regions [65]. The Project aimed to monitor trends in resistance, evaluate the performance of TB control programs, and advise on drug regimens. The Supranational TB Reference Laboratory Network (SRLN), a network now consisting of 29 laboratories globally, was developed in order to provide quality assurance to drug resistance surveys including panel testing before the start of a survey and rechecking of isolates during the survey [66]. Since 1994 drug resistance data have been systematically collected and analyzed from 119 countries worldwide (62% of all countries of the world). Out of them, 48 countries can rely on continuous surveillance systems based on routine diagnostic drug susceptibility testing of all patients. The remaining 71 countries have relied on special surveys of representative samples of patients.

The distributions of multidrug-resistant TB (MDR-TB), defined as TB caused by strains of *Mycobacterium tuberculosis* that are resistant to at least isoniazid and rifampicin, are given in Figs. 1 and 2, for new and



Tuberculosis, Epidemiology of. Figure 1
Distribution of MDR-TB among new TB cases, 1994–2010



Tuberculosis, Epidemiology of. Figure 2

Distribution of MDR-TB among previously treated TB cases, 1994–2010

previously treated TB cases, respectively. Proportions of MDR-TB exceeding 18% among new TB cases (in countries reporting more than ten MDR-TB cases) have been documented in Estonia (22.0%), Russian Federation (Arkhangelsk Oblast, 25.7%, Belgorod Oblast, 19.8%, Ivanovo Oblast, 20.3%, Kaliningrad Oblast, 22.3%, Murmansk Oblast, 28.9%, Pskov Oblast, 24.3%, and Vladimir Oblast, 20.9%). Proportions of MDR-TB exceeding 50% among previously treated TB cases (for countries reporting more than ten MDR-TB cases) are found in Estonia (51.6%), Lithuania (51.5%), Russian Federation (Arkhangelsk Oblast, 58.8%, Belgorod Oblast, 51.6%, Ivanovo Oblast, 57.7%, and Tomsk Oblast, 53.8%), and Tajikistan (Dushanbe city and Rudaki district, 61.6%, 95% CI: 52.5–70.2) [67, 68].

Since 2006, WHO has also been collecting and analyzing data on resistance to second-line anti-TB drugs. Extensively drug-resistant TB (XDR-TB), defined as MDR-TB plus resistance to a fluoroquinolone and at least 1 second-line injectable agent – amikacin, kanamycin, and/or capreomycin-, has been documented in 69 countries globally [67].

Overall, there were an estimated 390,000–510,000 cases of MDR-TB (primary and acquired) arising in 2008, with the best estimate at 440,000 cases. Among all incident TB cases globally, 3.6% (95% CI: 3.0–4.4) are estimated to have MDR-TB.

The estimated global number of incident MDR-TB episodes among new and relapse TB cases in 2008 was between 310,000 and 430,000 episodes, with the best estimate at 360,000 episodes. The estimated global number of incident acquired MDR-TB episodes was between 83,000 and 110,000 episodes, with the best estimate at 94,000 episodes. Almost 50% of MDR-TB cases worldwide are estimated to occur in China and India. These estimates refer to cases of MDR-TB that arose in 2008 and do not reflect the number of prevalent cases of MDR-TB. The number of prevalent cases of MDR-TB in many parts of the world is estimated to be much higher than the number arising annually.

An estimated 150,000 deaths caused by MDR-TB occurred globally in 2008, including those with HIV infection (range: 53,000–270,000). The estimated number of MDR-TB deaths excluding those with HIV infection was 97,000 (range: 6,000–220,000) [69].

Although the association of HIV and MDR-TB has been widely documented in hospital outbreaks of drug-resistant TB among people living with HIV, based on the population-based data gathered till now, it is not possible to conclude whether an overall association between MDR-TB and HIV epidemics exists [70–72].

More work is needed to understand global trends of the MDR-TB epidemic but in a group of countries and settings, decreasing trends in absolute numbers of

MDR-TB are documented (Estonia, Latvia, China, Hong Kong, United States of America, and two Oblasts in the Russian Federation, Tomsk and Orel) [69].

Control of Tuberculosis

In the second half of the nineteenth century more than one third of deaths globally were due to infectious diseases and 15% were due to tuberculosis. In particular, more than 30% of men and women during their productive years died of tuberculosis.

In industrialized nations, before the availability of the first anti-mycobacterial antibiotics in the second half of 1900, the incidence of tuberculosis notification dramatically decreased because of the reduction or elimination of some risk factors favoring the spread of mycobacterial strains in the community (for instance, improvement of the social and economic conditions of many families) and the implementation of public health interventions, such as the identification and isolation of contagious tuberculosis patients [1, 5, 8, 16, 73].

The first sanatorium was built in Germany in 1857 in order to isolate contagious tuberculosis patients and to provide them with adequate food, rest, sunlight, and fresh air. The dispensary system was introduced in 1897 in Scotland, England, following the basic German principle of the separation of an infectious patient from the community. After the Italian, Forlanini proved the efficacy of artificial pneumothorax to increase the likelihood of cure in tuberculosis cases in 1907, sanatoria increased their surgical activities to create pulmonary collapse (for instance, plombage and thoracoplasty) [1, 5, 8, 16, 74, 75].

The wide distribution of antituberculosis drugs and the scale-up of therapeutic combination, together with the establishment of the first National Tuberculosis Programs (NTPs) and the further improvement of the socio-economic situation, have been significantly associated with a rapid decline of tuberculosis incidence and prevalence in the USA and in numerous European countries [1, 5, 8, 16]. However, in the mid-1980s the erroneous feeling of imminent elimination of tuberculosis, supported by the persistent decline of incidence and prevalence, has led to a reduction of tuberculosis control measures with numerous sanatoria closed or converted into non-tuberculosis specialized facilities. The

escalation of the HIV/AIDS epidemic, the increasing number of migrants from high tuberculosis incidence countries, socioeconomic changes due to the collapse of the Soviet Union contributed to the rapid increase of tuberculosis incidence in low-incidence countries [1, 5, 8, 16, 76, 77]. Epidemiological trends in low-income countries are partially known but in the last decades the improvement of surveillance systems has increased the confidence on morbidity and mortality estimates [5, 8].

The World Health Organization has played a crucial role in the fight against tuberculosis worldwide. In 1993, tuberculosis was declared a global public health emergency and reiterated in 2000 by the United Nations Millennium Development Goals (MDGs). The Millennium Development Goal mainly focused on the tuberculosis issue aims at combating HIV, malaria, and other infectious diseases.

In the last 20 years, two public health strategies have changed the natural history of tuberculosis worldwide: DOTS strategy since 1996 and Stop TB Strategy since 2006. Both of them were launched after interactive and proactive debates and discussions between governmental and non-governmental organizations and national representatives of high tuberculosis incidence countries (tuberculosis control program managers and their staffs) [8, 78–82].

DOTS strategy is based on finding and treating contagious cases. It consists of five components: government commitment to tuberculosis control, bacteriological diagnosis through smear microscopy (mainly on individuals complaining of characteristic tuberculosis symptoms), short-course antituberculosis therapy that has to be standardized and supervised (i.e., directly observed, at least during the initial phase), uninterrupted high-quality drug supply, and individual outcome evaluation through a standardized recording and reporting system. Its targets were to detect 70% of sputum smear-positive tuberculosis cases and to successfully treat 85% of them by 2005. DOTS can reduce annual tuberculosis notifications from 6% to 8%, decreasing the incidence, the prevalence, and the mortality.

In 2005, the World Health Organization launched the new Stop TB Strategy (Table 2) [82]. The former strategy was revisited to pursue DOTS expansion, adding six other components to be implemented to

Tuberculosis, Epidemiology of. Table 2 WHO-recommended Stop TB Strategy [82]

1.	<i>Pursue high-quality DOTS expansion and enhancement</i> • Secure political commitment, with adequate and sustained financing • Ensure early case detection, and diagnosis through quality-assured bacteriology • Provide standardized treatment with supervision, and patient support • Ensure effective drug supply and management • Monitor and evaluate performance and impact
2.	<i>Address TB/HIV, MDR-TB, and the needs of poor and vulnerable populations</i> • Scale up collaborative TB/HIV activities • Scale up prevention and management of multidrug-resistant TB (MDR-TB) • Address the needs of TB contacts, and of poor and vulnerable populations, including women, children, prisoners, refugees, migrants, and ethnic minorities
3.	<i>Contribute to health system strengthening based on primary health care</i> • Help improve health policies, human resource development financing, supplies, service delivery, and information • Strengthen infection control in health services, other congregate settings, and households • Upgrade laboratory networks, and implement the Practical Approach to Lung Health (PAL) • Adapt successful approaches from other fields and sectors, and foster action on the social determinants of health
4.	<i>Engage all care providers</i> • Involve all public, voluntary, corporate, and private providers through Public–Private Mix (PPM) approaches • Promote use of the International Standards for TB Care (ISTC)
5.	<i>Empower people with TB, and communities through partnership</i> • Pursue advocacy, communication, and social mobilization • Foster community participation in TB care • Promote use of the Patients' Charter for TB care
6.	<i>Enable and promote research</i> • Conduct program-based operational research, and introduce new tools into practice • Advocate for and participate in research to develop new diagnostics, drugs, and vaccines

reach the 2015 Millennium Development Goal related to tuberculosis control, that is, to reduce prevalence of and mortality due to tuberculosis by 50% relative to those of 1990.

The new Stop TB Strategy can be summarized in the following elements [82]: (1) pursuing high-quality DOTS expansion and enhancement; (2) addressing tuberculosis/HIV coinfection, multi-drug resistant tuberculosis, and other challenges; (3) contributing to the strengthening of health-care systems; (4) engaging all care providers; (5) empowering people and communities with tuberculosis; and (6) enabling and promoting research.

This new Strategy addresses several issues that emerged during the implementation and scale-up of the DOTS strategy: spread of tuberculosis/HIV coinfection and increased rates of multi-drug resistant tuberculosis, weak health care systems, exclusion of the private sector from national tuberculosis control strategies (often managing poor and marginalized individuals), passive involvement of tuberculosis patients and communities, few resources in research and development of new diagnostics, drugs, and vaccines.

The first component, that is, “to pursue quality DOTS expansion and enhancement” is aimed at improving the previous strategy: political commitment should be strong in terms of financial resources available for tuberculosis control at national- and regional-level as well as for strengthening of health system. National laboratory networks should be improved with the introduction of culture methods and drug susceptibility testing as well as new effective evidence-based diagnostic technologies. In order to obtain high-treatment rates and to avoid the emergence and spread of drug-resistant strains treatment support should be provided through a patient-centered approach. A national body should manage the continuous supply of quality-assured anti-tuberculosis drugs. Continuous monitoring of the national TB program and evaluation of the tuberculosis control activities should be regularly implemented.

The second component of the strategy describes all the interventions that should be implemented to control and avoid further emergence of the rising issues: tuberculosis/HIV coinfection and multi-drug resistant tuberculosis. Tuberculosis and HIV national programs

should implement collaborative activities. Patients with drug-resistant tuberculosis should have access to diagnosis and proper treatment with second-line anti-TB drugs. More attention should be given to groups at highest risk of developing tuberculosis (such as prisoners, immigrants, etc.).

The third component is focused on strengthening of health system, highlighting all the elements necessary to improve a national health system: human and financial resources, information systems, and national policies.

The fourth component underlines the importance of the engagement of all health-care providers, including those outside national health system (public-private mix approach) and those working in prisons, general hospitals, and medical colleges.

The fifth component is aimed at describing the empowerment of people with tuberculosis and of communities, sharing some tuberculosis activities like treatment support. Furthermore, communities could have a role in reducing tuberculosis stigma.

The sixth component highlights the importance of the operational and biomedical research aimed at improving program performance and at developing new diagnostic, therapeutic, and preventive tools.

The Stop TB Strategy addresses all the risk factors associated to the development of mycobacterial infection and tuberculosis, including the reduction of the social and economic consequences of the disease. It supports the Global Plan to Stop TB (2011–2015), developed by the Stop TB Partnership [81]. The plan is aimed at identifying political and public health strategies to be implemented, and financial needs and existing gaps to address in order to scale up the Stop TB Strategy. Finally, International Standards of Tuberculosis Care (ISTC) have been issued to help both public and private health care providers to deliver quality DOTS services [83–85]. The aim is to describe an evidence-based standard of care for health-care workers managing confirmed or suspected tuberculosis individuals. The following principles of care are highlighted: early and accurate bacteriological diagnosis; administration of standardized and effective anti-tuberculosis drug regimens as well as appropriate treatment support and supervision; and monitoring of treatment outcomes.

Global Epidemiological Situation

Incidence

According to the last global estimates the number of new cases of tuberculosis in 2009 was 9.4 million, ranging from 8.9 million to 9.9 million (137 cases per 100,000 inhabitants). During the recent years the incidence trend has slowly decreased but the absolute frequency has increased due to population growth. The regions contributing the most to the global estimated incidence of tuberculosis are the South-east Asian region with 35% of cases (3,300,000), the Western Pacific region with 20% (1,900,000 cases), the African region with 30% (2,800,000 cases) followed by the Eastern Mediterranean region (7%), the European region (4%), and the Region of the Americas (3%) [4, 6].

The World Health Organization identified 22 so-called high burden countries (HBCs), accounting for more than 80% of the incident cases estimated globally. India accounts for 21% of all tuberculosis patients and China for 14%. The highest incidence rates in 2009 were estimated in India (range: 1.6–2.4 million), China (range: 1.1–1.5 million), South Africa (range: 0.40–0.59 million), Nigeria (range: 0.37–0.55 million), and Indonesia (range: 0.35–0.52 million). The incident tuberculosis/HIV coinfection was estimated in 1.1/9.4 million (12%; 1.0–1.2 million, 11–13%) individuals, with the vast majority of patients in the African region. Among the incident cases estimated in 2008, 440,000 (range: 390,000–510,000) patients were infected by multi-drug resistant mycobacterial strains, that is, with resistance *in vitro* to at least isoniazid and rifampicin (MDR-TB).

Prevalence

The estimated tuberculosis prevalence in 2009 was 200 cases per 100,000 inhabitants as a consequence of 14 million (range: 12–16 million) prevalent patients.

Mortality

In 2009 about 1.7 million individuals (i.e., 26 cases per 100,000 inhabitants) died of tuberculosis, with 400,000 (range: 320,000–450,000) and 1.3 million (range: 1.2–1.5 million) deaths among tuberculosis/HIV coinfecting and HIV-negative patients, respectively.

Case Notifications

National tuberculosis programs registered 5.8 million tuberculosis patients in 2009: most of them were new cases (94.8%) and 0.3 million were relapses (i.e., previously treated cases whose most recent anti-tuberculosis treatment outcome was “cure” or “treatment completed” but who were afterward diagnosed with sputum smear and/or culture-positive tuberculosis); 57% of the pulmonary tuberculosis cases was considered highly infectious, that is, sputum smear-positive. Among the 5.5 million new cases notified in 2009, 47.3% (2.6 million) was sputum smear-positive, 36.4% (2.0 million) sputum smear-negative while 16.4% (900,000) was affected by extrapulmonary tuberculosis.

The causes behind the global under-reporting could be linked to the patients (limited knowledge of tuberculosis; economic and/or geographical barriers to access health care; etc.) and/or to the health systems (poor clinical and laboratory capacities; few trained health care workers; poor compliance to national laws on notification after tuberculosis diagnosis).

Treatment Outcomes

The global treatment success (“cured” plus “treatment completed”) rates for new sputum smear-positive cases treated during the years 2007 and 2008 were higher than the target set in 1991, that is, 85%.

In 2008 three WHO regions contributed to this result: Eastern Mediterranean Region, the South-east Asian Region, and the Western Pacific Region; the worst treatment success rate was documented in the European Region (66%), followed by the Region of the Americas (77%) and the African Region (80%).

Future Directions

Elimination of tuberculosis, or reducing the incidence of new sputum smear-positive cases below one per one million inhabitants, is the goal for the next decades. The target endorsed by Stop TB Partnership is the elimination of tuberculosis as a public health problem by 2050 [6–88].

The objectives of the tuberculosis elimination strategy are:

- Reduction of the incidence of latent tuberculosis infection.
- Reduction of the prevalence of latent tuberculosis infection.

The aim is to decrease the burden of future new tuberculosis cases.

In low tuberculosis-incidence countries, the lowest prevalence and incidence of latent tuberculosis infection are detected in the autochthonous youngest groups. The epidemiological projections indicate that each generation will be substituted by a generation with a lower risk of infection and, then, characterized by a lower prevalence of latent tuberculosis infections [5, 8, 86, 88].

Early detection and adequate treatment of contagious tuberculosis patients are of paramount importance in order to achieve a positive epidemiological trend. Moreover, the progression to tuberculosis in individuals infected by mycobacterial strains should be prevented, particularly focusing on high-risk groups (migrants, HIV-positives, ethnic minorities, prisoners, elderly people, and close contacts of tuberculosis index cases) [5, 8, 86].

This combined approach is deemed essential to reach the elimination goal, and it should be supported by international and national policies, embracing the following components [2, 5, 8, 86, 87]:

- Political commitment toward control and elimination, particularly providing human resources, facilities, and funds
- Case detection through case-finding among symptomatic individuals presenting at health services and active case finding in special groups through quality-assured laboratories and trained health-care providers
- Access to tuberculosis diagnostic and treatment services, implementing collaborative activities with health bodies or organizations working with migrants, prisoners, homelessness, etc.
- Standard approach to treatment of tuberculosis and latent tuberculosis infection, following international evidence-based recommendations
- Surveillance and treatment outcome monitoring, following standard internationally accepted definitions

However, tuberculosis indicators could potentially worsen in the next decades due to the following public health challenges:

- Drug-resistant mycobacterial strains (mostly multi/extensively drug-resistant mycobacteria)
- HIV/AIDS
- Immunocompromised individuals
- Political unrests and wars, natural disasters, and starvation

Adequate funds, constant international and national political commitment, and an effective supply of anti-tuberculosis and anti-HIV drugs and diagnostics represent relevant components for a successful strategy. In particular, the research and development of new effective anti-tuberculosis drugs with an improved toxicity, tolerability, and efficacy profile and the evaluation of the activity of potential adjunctive treatments, such as vitamin D, as well as of new primary preventive measures (i.e., vaccines) are crucial priorities for the near future. Basic and applied research would need to be followed by operational research, as new diagnostic, therapeutic, and preventive approaches must be tested at the program-level (field trials) after clinical trials [89].

Bibliography

1. Fitzgerald D, Haas DW (2005) *Mycobacterium tuberculosis*. In: Mandell GL et al (eds) Principles and practice of infectious diseases, 6th edn. Churchill Livingstone, Philadelphia, pp 2852–2886
2. Nunn P, Williams B, Floyd K, Dye C, Elzinga G, Raviglione M (2005) Tuberculosis control in the era of HIV. *Nat Rev Immunol* 5(10):819–826
3. Raviglione MC (2006) The global plan to stop TB, 2006–2015. *Int J Tuberc Lung Dis* 10(3):238–239
4. World Health Organization (2010) Global tuberculosis control. A short update to the 2009 report. WHO, Geneva, 2010
5. Rieder H (1999) Epidemiologic basis of tuberculosis control. International Union Against Tuberculosis and Lung Disease, Paris
6. World Health Organization (2010) Global tuberculosis control 2010. World Health Organization Document 2010, Publication No. WHO/HTM/TB/2010.7
7. Everitt BS, Palmer C (2011) Encyclopedic companion to medical statistics, 2nd edn. Wiley, Chichester
8. Styblo K, Raviglione MC (1997) Tuberculosis, public health aspects. In: Encyclopedia of human biology, vol 8, 2nd edn. Academic, San Diego, pp 537–558
9. Styblo K, Danková D, Drápela J, Galliová J, Jezek Z, Krivánek J, Kubík A, Langerová M, Radkovský J (1967) Epidemiological and clinical study of tuberculosis in the district of Kolin, Czechoslovakia. Report for the first 4 years of the study (1961–64). *Bull World Health Organ* 37(6):819–874
10. Krivinka R, Drápela J, Kubík A, Danková D, Krivánek J, Ruzha J, Miková Z, Hejdrová E (1974) Epidemiological and clinical study of tuberculosis in the district of Kolín, Czechoslovakia. Second report (1965–1972). *Bull World Health Organ* 51(1):59–69
11. Borgdorff MW, Nagelkerke NJ, van Soolingen D, Broekmans JF (1999) Transmission of tuberculosis between people of different ages in The Netherlands: an analysis using DNA fingerprinting. *Int J Tuberc Lung Dis* 3(3):202–206
12. Rieder HL (1999) Socialization patterns are key to the transmission dynamics of tuberculosis. *Int J Tuberc Lung Dis* 3(3):177–178
13. Stead WW (1998) Tuberculosis among elderly persons, as observed among nursing home residents. *Int J Tuberc Lung Dis* 2(9 Suppl 1):S64–S70
14. Leung CC, Rieder HL, Lange C, Yew WW (2011) Treatment of latent infection with *Mycobacterium tuberculosis*: update 2010. *Eur Respir J* 37(3):690–711
15. WHO (2009) WHO policy on infection control in health-care facilities, congregate settings and households. World Health Organization, Geneva. WHO/HTM/TB/2009.419
16. Migliori GB, Sotgiu G, Lange C, Centis R (2010) Extensively drug-resistant tuberculosis: back to the future. *Eur Respir J* 36(3):475–477
17. Storla DG, Yimer S, Bjune GA (2008) A systematic review of delay in the diagnosis and treatment of tuberculosis. *BMC Public Health* 8:15
18. Ngadaya ES, Mfinanga GS, Wandwalo ER, Morkve O (2009) Delay in tuberculosis case detection in Pwani region, Tanzania. A cross sectional study. *BMC Health Serv Res* 9:196
19. European Centers of Disease Prevention and Control (2011) Use of interferon- γ release assays in support of TB diagnosis. ECDC, Stockholm, 2011
20. Cellestis (2006) QuantiFERON-TB Gold (In-Tube Method) Package Insert. Valencia, 2006
21. Oxford Immunotec (2009) T-Spot. TB Package Insert. 2009
22. Diel R, Goletti D, Ferrara G, Bothamley G, Cirillo D, Kampmann B, Lange C, Losi M, Markova R, Migliori GB, Nienhaus A, Ruhwald M, Wagner D, Zellweger JP, Huitric E, Sandgren A, Manissero D (2011) Interferon- γ release assays for the diagnosis of latent *Mycobacterium tuberculosis* infection: a systematic review and meta-analysis. *Eur Respir J* 37(1):88–99
23. Torrado E, Robinson RT, Cooper AM (2011) Cellular response to mycobacteria: balancing protection and pathology. *Trends Immunol* 32(2):66–72
24. Creswell J, Raviglione M, Ottmani S, Migliori GB, Uplekar M, Blanc L, Sotgiu G, Lonnroth K (2010) Tuberculosis and non-communicable diseases: neglected links, missed opportunities. *Eur Respir J* 37(5):1269–1282

25. Lönnroth K, Jaramillo E, Williams BG, Dye C, Raviglione M (2009) Drivers of tuberculosis epidemics: the role of risk factors and social determinants. *Soc Sci Med* 68(12):2240–2246
26. Schutz C, Meintjes G, Almajid F, Wilkinson RJ, Pozniak A (2010) Clinical management of tuberculosis and HIV-1 co-infection. *Eur Respir J* 36(6):1460–1481
27. World Health Organization (2006) Diabetes fact sheet No. 312
28. Jeon CY, Murray MB (2008) Diabetes mellitus increases the risk of active tuberculosis: a systematic review of 13 observational studies. *PLoS Med* 5(7):e152
29. Stevenson CR, Forouhi NG, Roglic G, Williams BG, Lauer JA, Dye C, Unwin N (2007) Diabetes and tuberculosis: the impact of the diabetes epidemic on tuberculosis incidence. *BMC Public Health* 7:234
30. Stevenson CR, Critchley JA, Forouhi NG, Roglic G, Williams BG, Dye C, Unwin NC (2007) Diabetes and the risk of tuberculosis: a neglected threat to public health? *Chronic Illn* 3(3):228–245
31. Harries AD, Billo N, Kapur A (2009) Links between diabetes mellitus and tuberculosis: should we integrate screening and care? *Trans R Soc Trop Med Hyg* 103(1):1–2
32. Dooley KE, Chaisson RE (2009) Tuberculosis and diabetes mellitus: convergence of two epidemics. *Lancet Infect Dis* 9(12):737–746
33. Martens GW, Arian MC, Lee J, Ren F, Greiner D, Kornfeld H (2007) Tuberculosis susceptibility of diabetic mice. *Am J Respir Cell Mol Biol* 37(5):518–524
34. Stalenhoef JE, Alisjahbana B, Nelwan EJ, van der Ven-Jongekrijg J, Ottenhoff TH, van der Meer JW, Nelwan RH, Netea MG, van Crevel R (2008) The role of interferon-gamma in the increased tuberculosis risk in type 2 diabetes mellitus. *Eur J Clin Microbiol Infect Dis* 27(2):97–103
35. Lienhardt C, Rodrigues LC (1997) Estimation of the impact of the human immunodeficiency virus infection on tuberculosis: tuberculosis risks re-visited? *Int J Tuberc Lung Dis* 1(3):196–204
36. Hussein MM, Mooij JM, Roujouleh H (2003) Tuberculosis and chronic renal disease. *Semin Dial* 16(1):38–44
37. Woeltje KF, Mathew A, Rothstein M, Seiler S, Fraser VJ (1998) Tuberculosis infection and anergy in hemodialysis patients. *Am J Kidney Dis* 31(5):848–852
38. Girndt M, Sester U, Sester M, Kaul H, Köhler H (1999) Impaired cellular immune function in patients with end-stage renal failure. *Nephrol Dial Transplant* 14(12):2807–2810
39. Rees D, Murray J (2007) Silica, silicosis and tuberculosis. *Int J Tuberc Lung Dis* 11(5):474–484
40. Barboza CE, Winter DH, Seiscento M, Santos Ude P, Terra Filho M (2008) Tuberculosis and silicosis: epidemiology, diagnosis and chemoprophylaxis. *J Bras Pneumol* 34(11):959–966
41. Hnizdo E, Murray J (1998) Risk of pulmonary tuberculosis relative to silicosis and exposure to silica dust in South African gold miners. *Occup Environ Med* 55(7):496–502
42. Bates MN, Khalakdina A, Pai M, Chang L, Lessa F, Smith KR (2007) Risk of tuberculosis from exposure to tobacco smoke: a systematic review and meta-analysis. *Arch Intern Med* 167(4):335–342
43. Arcavi L, Benowitz NL (2004) Cigarette smoking and infection. *Arch Intern Med* 164(20):2206–2216
44. Solovic I, Sester M, Gomez-Reino JJ, Rieder HL, Ehlers S, Milburn HJ, Kampmann B, Hellmich B, Groves R, Schreiber S, Wallis RS, Sotgiu G, Schölvinc EH, Goletti D, Zellweger JP, Diel R, Carmona L, Bartalesi F, Ravn P, Bossink A, Duarte R, Erkens C, Clark J, Migliori GB, Lange C (2010) The risk of tuberculosis related to tumour necrosis factor antagonist therapies: a TBNET consensus statement. *Eur Respir J* 36(5):1185–1206
45. Ehlers S (2005) Tumor necrosis factor and its blockade in granulomatous infections: differential modes of action of infliximab and etanercept? *Clin Infect Dis* 41(Suppl 3):S199–S203
46. Bekker LG, Freeman S, Murray PJ, Ryffel B, Kaplan G (2001) TNF-alpha controls intracellular mycobacterial growth by both inducible nitric oxide synthase-dependent and inducible nitric oxide synthase-independent pathways. *J Immunol* 166(11):6728–6734
47. Raviglione MC, Harries AD, Msiska R, Wilkinson D, Nunn P (1997) Tuberculosis and HIV: current status in Africa. *AIDS* 11(Suppl B):S115–S123
48. Getahun H, Gunneberg C, Granich R, Nunn P (2010) HIV infection-associated tuberculosis: the epidemiology and the response. *Clin Infect Dis* 50(Suppl 3):S201–S207
49. Joint United Nations Programme on HIV/AIDS (2008) Report on the global AIDS epidemic. Joint United Nations Programme on HIV/AIDS, Geneva, 2008
50. Corbett EL, Watt CJ, Walker N, Maher D, Williams BG, Raviglione MC, Dye C (2003) The growing burden of tuberculosis: global trends and interactions with the HIV epidemic. *Arch Intern Med* 163(9):1009–1021
51. Reid A, Scano F, Getahun H, Williams B, Dye C, Nunn P, De Cock KM, Hankins C, Miller B, Castro KG, Raviglione MC (2006) Towards universal access to HIV prevention, treatment, care, and support: the role of tuberculosis/HIV collaboration. *Lancet Infect Dis* 6(8):483–495
52. World Health Organisation (2006) Antiretroviral therapy for HIV infection in adults and adolescents in resource limited settings: towards universal access. World Health Organisation, Geneva
53. World Health Organization (2010) Treatment of tuberculosis guidelines, 4th edn. World Health Organization, Geneva, Document WHO/HTM/TB/2009.420
54. Abdool Karim S, Naidoo K, Grobler A, et al. Initiating ART during TB treatment significantly increases survival: results of a randomised controlled clinical trial in TB/HIV co-infected patients in South Africa. Presented at 16th conference on retroviruses and opportunistic infections. Montreal, 8–11 Feb 2009
55. WHO (2011) Guidelines for intensified tuberculosis case-finding and isoniazid preventive therapy for people living with HIV in resource constrained settings. WHO, Geneva, 2011
56. WHO Three I's Meeting. 2008

57. Medical Research Council (1948) Streptomycin treatment of pulmonary tuberculosis. A Medical Research Council investigation. *BMJ* 2:769–782
58. Fox W, Wiener A, Mitchison DA, Selkon JB, Sutherland I (1957) The prevalence of drug-resistant tubercle bacilli in untreated patients with pulmonary tuberculosis: a national survey, 1955–1956. *Tubercle* 38:71
59. Zhang Y, Yew WW (2009) Mechanisms of drug resistance in *Mycobacterium tuberculosis*. *Int J Tuberc Lung Dis* 13(11):1320–1330
60. Wright A, Zignol M, Van Deun A, Falzon D, Gerdes SR, Feldman K, Hoffner S, Drobniowski F, Barrera L, van Soolingen D, Boulabhal F, Paramasivan C, Kam KM, Mitarai S, Nunn P, Raviglione M (2009) For the Global Project on Anti-Tuberculosis Drug Resistance Surveillance. Epidemiology of antituberculosis drug resistance 2002–07: an updated analysis of the Global Project on Anti-Tuberculosis Drug Resistance Surveillance. *Lancet* 373(9678):1861–1873
61. Ellner JJ, Hinman AR, Dooley SW et al (1993) Tuberculosis symposium: emerging problems and promise. *J Infect Dis* 168:537–551
62. Frieden TR, Sterling T, Pablos-Mendez A et al (1993) The emergence of drug-resistant tuberculosis in New York City. *N Engl J Med* 328:521–526
63. Rastogi N (1993) Emergence of multiple-drug-resistant tuberculosis: fundamental and applied research aspects, global issues and current strategies. *Res Microbiol* 144:103
64. Sbarbaro JA (1993) TB control in the 21st century. *Monaldi Arch Chest Dis* 48:197–198, Editorial
65. Cohn DL, Bustreo F, Raviglione MC (1997) Drug-resistant tuberculosis: review of the worldwide situation and the WHO/IUATLD global surveillance project. *Clin Infect Dis* 24(Suppl 1):S121–S130
66. Van Deun A, Wright A, Zignol M, Weyer K, Rieder HL (2011) Drug susceptibility testing proficiency in the network of supranational tuberculosis reference laboratories. *Int J Tuberc Lung Dis* 15(1):116–124
67. World Health Organization (2011) Towards universal access to diagnosis and treatment of multidrug-resistant and extensively drug-resistant tuberculosis by 2015 (WHO/HTM/TB/2011.3). Geneva, 2011
68. Ministry of Health and Social Development of the Russian Federation (2010) Tuberculosis in the Russian Federation 2009. An analytical review of the TB statistical indicators used in the Russian Federation. The Russian Federation, Moscow, 2010
69. World Health Organization (2010) Multidrug and extensively drug-resistant TB (M/XDR-TB) - 2010 global report on surveillance and response (WHO/HTM/TB/2010.3). Geneva, 2010
70. Edlin BR, Tokars JI, Grieco MH, Crawford JT, Williams J, Sordillo EM, Ong KR, Kilburn JO, Dooley SW, Castro KG, Jarvis WR, Holmberg SD (1992) An outbreak of multidrug-resistant tuberculosis among hospitalized patients with the acquired immunodeficiency syndrome. *N Engl J Med* 326:1514–1521
71. Moro ML, Gori A, Errante I, Infuso A, Franzetti F, Sodano L, Iemoli E (1998) An outbreak of multidrug-resistant tuberculosis involving HIV-infected patients of two hospitals in Milan, Italy. Italian Multidrug-Resistant Tuberculosis Outbreak Study Group. *AIDS* 12:1095–1102
72. Wells CD, Cegielski JP, Nelson LJ, Laserson KF, Holtz TH, Finlay A, Castro KG, Weyer K (2007) HIV infection and multidrug-resistant tuberculosis: the perfect storm. *J Infect Dis* 196(Suppl 1):S86–S107
73. Raviglione M (2006) XDR-TB: entering the post-antibiotic era? *Int J Tuberc Lung Dis* 10(11):1185–1187
74. Dawson JJ, Devadatta S, Fox W, Radhakrishna S, Ramakrishnan CV, Somasundaram PR et al (1966) A 5-year study of patients with pulmonary tuberculosis in a concurrent comparison of home and sanatorium treatment for one year with isoniazid plus PAS. *Bull World Health Organ* 34(4):533–551
75. Styblo K, Meijer J, Sutherland I (1969) Tuberculosis surveillance research unit report no. 1: the transmission of tubercle bacilli; its trend in a human population. *Bull Int Union Tuberc* 42: 1–104
76. Dahle UR, Eldholm V, Winje BA, Mannsåker T, Haldal E (2007) Impact of immigration on the molecular epidemiology of *Mycobacterium tuberculosis* in a low-incidence country. *Am J Respir Crit Care Med* 176(9):930–935
77. McKenna MT, McCray E, Onorato I (1995) The epidemiology of tuberculosis among foreign-born persons in the United States, 1986 to 1993. *N Engl J Med* 332(16):1071–1076
78. World Health Organisation (2015) The global plan to stop TB 2006–2015
79. World Health Organization (2005) Fifty-eighth world health assembly: resolutions and decisions. WHA58/2005/REC/1. WHO, Geneva, 2005
80. WHO (2004) Report on the meeting of the second ad hoc Committee on the TB Epidemic (WHO/HTM/STB/2004.28). World Health Organization, Geneva, 2004
81. WHO (2006) Stop TB partnership and the World Health Organization. The global plan to stop tuberculosis, 2006–2015. WHO/HTM/STB/2006.35. WHO, Geneva, 2006
82. Raviglione MC, Uplekar MW (2006) WHO's new stop TB strategy. *Lancet* 367(9514):952–955
83. Migliori GB, Hopewell PC, Blasi F, Spanevello A, Raviglione MC (2006) Improving the TB case management: the international standards for tuberculosis care. *Eur Respir J* 28(4):687–690
84. Tuberculosis Coalition for Technical Assistance (2006) International standards for tuberculosis care (ISTC). Tuberculosis Coalition for Technical Assistance, The Hague, 2006
85. Tuberculosis Coalition for Technical Assistance (2009) International standards for tuberculosis care (ISTC), 2nd edn. Tuberculosis Coalition for Technical Assistance, The Hague, 2009
86. Clancy L, Rieder HL, Enarson DA, Spinaci S (1991) Tuberculosis elimination in the countries of Europe and other industrialized countries. *Eur Respir J* 4(10):1288–1295
87. Broekmans JF, Migliori GB, Rieder HL, Lees J, Ruutu P, Løddenkemper R, Raviglione MC (2002) World Health

- Organization, International Union Against Tuberculosis and Lung Disease, and Royal Netherlands Tuberculosis Association Working Group. European framework for tuberculosis control and elimination in countries with a low incidence. Recommendations of the World Health Organization (WHO), International Union Against Tuberculosis and Lung Disease (IUATLD) and Royal Netherlands Tuberculosis Association (KNCV) Working Group. *Eur Respir J* 19(4):765–775
88. Migliori GB, Loddenkemper R, Blasi F, Raviglione MC (2007) 125 years after Robert Koch's discovery of the tubercle bacillus: the new XDR-TB threat. Is "science" enough to tackle the epidemic? *Eur Respir J* 29(3):423–427
89. Migliori GB, D'Arcy Richardson M, Sotgiu G, Lange C (2009) Multidrug-resistant and extensively drug-resistant tuberculosis in the West. Europe and United States: epidemiology, surveillance, and control. *Clin Chest Med* 30(4):637–665